НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 336.77.067.31

До захисту допущено
В. о. завідувача кафедри ММСА
О.Л.Тимощук
«___» _____ 2019 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 124 Системний аналіз
на тему «Регресійні методи для оцінювання вартості
портфеля кредитної заборгованості»

Виконав студент групи КА-81 мнв
Мацагор Іван Дмитрович _____

Керівник: доцент кафедри ММСА,
к.ф.-м.н, доц. Каніовська Ірина Юріївна. _____

Рецензент доцент кафедри математичного
аналізу та теорії ймовірностей
КПІ ім. Ігоря Сікорського
к.ф.-м.н., доц. Буценко Юрій Павлович _____

Засвідчую, що у цій магістерській
дисертації немає запозичень з праць інших
авторів без відповідних посилань
Студент _____

Київ
2019

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність — 124 «Системний аналіз»

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри ММСА
Тимощук О.Л.
«___» _____ 2019 р.

ЗАВДАННЯ
на магістерську дисертацію студенту Мацагору Івану Дмитровичу

1. Тема дисертації: «Регресійні методи для оцінювання вартості портфеля кредитної заборгованості», науковий керівник дисертації Каніовська Ірина Юріївна, к. ф.-м. н., доцент, затверджені наказом по університету від «___» __ _____ № _____

2. Термін подання студентом дисертації: 13 грудня 2019 р.

3. Об'єкт дослідження: модель оцінювання вартості портфеля кредитної заборгованості на основі регресійних методів.

4. Предмет дослідження: побудова та порівняння різних регресійних моделей для оцінювання вартості портфеля кредитної заборгованості.

5. Перелік завдань, які потрібно розробити:

1) дослідити сучасний стан та особливості застосування математичного моделювання у вирішенні проблеми проблемних кредитів;
2) розробити математичні моделі оцінювання вартості портфеля кредитної заборгованості за допомогою регресійних методів;
3) підібрати та обробити вибірку для навчання та тестування моделей;
4) порівняти точність різних моделей та вибрати найкращу;
5) розробити стартап-проект виведення на ринок результатів дослідження;

6) розробити концептуальні висновки за результатами наукового дослідження.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

1) точкові діаграми залежностей залежної величини від незалежних (рис.);

2) графіки гіпотетичних залежностей залежної величини від незалежних (рис.);

3) виводи терміналу R (рис.);

4) таблиці у розділі стартап-проекту.

7. Дата видачі завдання: 05 вересня 2019 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації
1.	Концептуальний вступ дисертації. Формулювання об’єкта, предмета, цілі, завдань, новизни, практичної значущості результатів	18.03.2019 – 20.03.2019
2.	Перший розділ. Огляд літературно-інформаційних джерел. Понятійно-категоріальний апарат. Характеристика об’єкта	21.03.2019 – 30.03.2019
3.	Другий розділ. Дослідження сімейства регресійних методів	31.03.2019 – 16.04.2019
4.	Третій розділ. Розробка моделі для оцінювання вартості портфеля кредитної заборгованості. Реалізація за допомогою засобів мови R	17.04.2019 – 02.05.2019
5.	Четвертий розділ. Стартап-проект	03.05.2019 – 06.05.2019
6.	Концептуальні висновки. Перспективи розвитку отриманих рішень	07.05.2019 – 10.05.2019

Студент І.Д.Мацагор

Науковий керівник дисертації І.Ю.Каніовська

РЕФЕРАТ

Магістерська дисертація: 96 с., 4 частини, 16 рис., 30 табл., 25 джерел.

Об’єкт дослідження: модель оцінювання вартості портфеля кредитної заборгованості на основі регресійних методів.

Предмет дослідження: побудова та порівняння різних регресійних моделей для оцінювання вартості портфеля кредитної заборгованості.

Метою даної магістерської дисертації є розробка робочої моделі для оцінювання вартості портфеля кредитної заборгованості, що дозволить банкам та колекторським компаніям України робити економічно обґрунтовані та взаємовигідні цінові пропозиції.

У роботі проведено огляд методів оцінювання вартості портфеля кредитної заборгованості в Україні; досліджені усі ключові фактори, що впливають на платоспроможність боржника; побудовано дві альтернативні моделі для оцінювання вартості портфеля кредитної заборгованості; проведено аналіз результатів для визначення кращої моделі; наведено висновки та пропозиції щодо можливих покращень моделі.

Результати та їх новизна: виявлено застарілість та недосконалість методів оцінювання вартості портфеля кредитної заборгованості у банках України; розробка моделі здійснювалась на основі реальної актуальної вибірки кредитних справ враховуючи сучасні українські реалії (серед факторів визначення платоспроможності боржника є приналежність його місця проживання до зони АТО).

РЕГРЕСІЙНІ МЕТОДИ, ЛОГІСТИЧНА РЕГРЕСІЯ, КРЕДИТНИЙ ПОРТФЕЛЬ, ПЛАТОСПРОМОЖНІСТЬ, КОЛЕКТОРСЬКА ДІЯЛЬНІСТЬ.

ABSTRACT

Master’s thesis: 96 pp., 4 parts, 16 figures, 30 tables, 25 sources.

Object of research: model for cost evaluation of debt portfolio based on regression methods.

Subject of research: development and comparison of different regression models for cost evaluation of debt portfolio.

The purpose of this master’s thesis is to build a working model for cost evaluation of debt portfolio which will allow banks and collecting agencies of Ukraine to make economically sound and mutually beneficial price offers.

This thesis contains review of methods of cost evaluation of debt portfolio in Ukraine; examination of all key factors which affect debtor’s solvency; development of two alternative models for cost evaluation of debt portfolio; results analysis to determine the best model; conclusions and possible improvements to the model.

Results and their novelty: the obsolete and imperfect methods of cost evaluation of debt portfolio of banks in Ukraine are revealed; the development of the model was conducted using real actual data set of credits considering existent Ukrainian realities (one of the factors which affect debtor’s solvency is whether or not his residence situated in the antiterrorist operation zone).

REGRESSION METHODS, LOGISTIC REGRESSION, DEBT PORTFOLIO, SOLVENCY, COLLECTION.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ 8

ВСТУП 9

1. ДОСЛІДЖЕННЯ КРЕДИТНОЇ СФЕРИ 13

1.1. Основні поняття 13

1.2. Визначення поняття проблемної заборгованості 15

1.3. Огляд існуючих методів оцінювання кредитних портфелів 17

1.3.1. Рамкова методика оцінки проблемних активів 17

1.3.2. Оцінка за допомогою нейронних мереж 23

1.3. Висновку до розділу 1 31

2. РІЗНОВИДИ ЛІНІЙНОЇ РЕГРЕСІЇ 32

2.1. Класична регресія 32

2.1.1. Означення регресії 32

2.1.2. Метод найменших квадратів (МНК) 33

2.1.3. Статистичні властивості МНК-оцінок 36

2.2. Узагальнена лінійна регресія 37

2.2.1. Сімейство експоненційних розподілів 38

2.2.2. Представники експоненційного сімейства 40

2.2.3. Метод максимальної правдоподібності 42

2.2.4. Використання методу УЛР 46

2.3. Логістична регресія та ROC-аналіз 48

2.3.1. Опис методу логістичної регресії 48

2.3.2. Основи ROC-аналізу 49

2.3.3. ROC-крива 52
2.3.4. Показник AUC 54
2.3.5. Визначення порогу відсікання 55
2.4. Висновки до розділу 2 56
3. АНАЛІЗ ДАНИХ ТА ПОБУДОВА МОДЕЛІ 58
3.1. Аналіз вибірки та попередня обробка даних 58
3.2. Побудова множинної регресії 67
3.3. Побудова альтернативної моделі 70
3.4. Порівняння моделей 74
3.5. Висновки до розділу 3 75
4. СТАРТАП-ПРОЕКТ «СМАРТ ДЕБТ» 77
4.1. Опис ідеї проекту 77
4.2. Технологічний аудит ідеї проекту 78
4.3. Аналіз ринкових можливостей запуску стартап-проекту 79
4.4. Розроблення ринкової стратегії проекту 87
4.5. Розроблення маркетингової програми стартап-проекту 89
4.6. Висновки до розділу 4 92
ВИСНОВКИ 93
ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ 97
ДОДАТОК А. ЛІСТИНГ ПРОГРАМИ 100

ПЕРЕЛІК СКОРОЧЕНЬ

AUC – Area Under Curve; площа під ROC-кривою

DPD – Days Past Due; кількість днів прострочення платежу

ROC – Receiver Operating Characteristic; робоча характеристика приймача

АТО – АнтиТерористична Операція

ІПН – Індивідуальний Податковий Номер

МНК – Метод Найменших Квадратів

ПІБ – Прізвище, Ім’я, по-Батькові

РНОКПП – Реєстраційний Номер Облікової Картки Платника Податків

УЛР – Узагальнена Лінійна Регресія

ВСТУП

Після кризи 2014 – 2015 років українська банківська система виявилася в безпрецедентній ситуації. Конфлікт на сході України, невирішена ситуація з Автономною Республікою Крим, а також спад економічної активності і девальвація гривні більш ніж утричі дуже негативно вплинули на доходи клієнтів банків. Зокрема була зроблена заява, що станом на липень 2017 року в Україні зафіксована найбільш висока частка непрацюючих кредитів за всю історію світових спостережень [1].

Згідно зі статистикою НБУ, свого піку обсяг проблемних активів у банківській системі досяг у липні 2017 року – тоді частка кредитів, що не обслуговуються більше трьох місяців і лежать мертвим вантажем, склала 591,53 млрд грн, або 58% від загального обсягу. І хоча з серпня по грудень 2017 року цей показник знизився до 580,5 млрд грн (54,9%), частка “поганих” активів в системі як і раніше домінує.

Однак не лише воєнний конфлікт із Росією став причиною цього антирекорду. Становище на кредитному ринку було далеко не ідеальним ще задовго до першого вторгнення. Так, наприклад, кишенькові банки фінансували бізнес своїх власників за кошти платників податків. Нерідко такі банки видавали кредити пов’язаним компаніям на неринкових умовах, і повертати їх від самого початку ніхто і не збирався. Після 2013 року більшість таких банків, які "пилососили" ринок депозитів, НБУ встиг вивести з ринку. Проте, проблеми з верховенством права та неспроможність судової системи захистити права кредиторів продовжують негативно впливати на якість активів і залишаються основним стримуючим фактором розвитку нового корпоративного кредитування.

Нинішнє законодавство дозволяє несумлінним позичальникам розтягувати судові процеси зі стягнення заборгованості на кілька років. Коли

банк отримує рішення суду про стягнення, боржник ініціює затяжну процедуру банкрутства і при цьому продовжує вести бізнес або використовувати свої активи, перевівши діяльність на іншу юридичну особу.

Не виправдалися надії банкірів і на Закон "Про фінансову реструктуризацію", що діє з жовтня 2017 року. Адже можливість часткового прощення заборгованості і податкові пільги виявилися не дуже цікавими для несумлінних позичальників в порівнянні з можливістю і далі безкарно нічого нікому не платити.

Якби банки відновили активне кредитування, частка токсичних активів в системі почала б поступово розмиватися за рахунок нових працюючих позик. Але цей процес поки йде мляво і навряд чи пришвидшиться після того, як Нацбанк наприкінці січня 2018 року підвищив облікову ставку відразу на 1,5 процентних пункта – до 16%. НБУ пішов на цей крок заради приборкання інфляції. Однак подібний вимушений захід може спричинити подорожчання кредитів, рецесію економіки і консервацію поточної ситуації з поганими активами.

Також слід зазначити, що основну частку “поганих” активів складають корпоративні (85%), а не споживчі кредити (на смартфони, телевізори тощо). Однак у цій роботі будуть розглядатися лише кредити, видані фізичним особам, адже вони є більш однорідними і набагато краще піддаються статистичній обробці, на відміну від корпоративних кредитів, для роботи з якими на даному етапі через недостатність досвіду необхідний індивідуальний аналіз кожного окремого кредиту [3].

У банківських закладах виділяють чотири групи неплатників. До перших двох відносяться ті, хто забув про свої зобов’язання, або тимчасово не може їх виконувати через погіршення фінансового стану. Працювати з такими позичальниками найпростіше – їм досить нагадування або короткого роз’яснення про наслідки невчасних розрахунків. Проблемною категорією є позичальники, що сподіваються, що банки про них забудуть або неплатіж

зійде їм з рук. Частіше всього такі позичальники не знають про відповідальність за прострочені борги, загубили або пошкодили придбані в кредит товари або недооцінили вартість кредиту. Як правило, боржники цієї групи відмовляються вносити платежі за кредитом у середині терміну погашення. При залученні до роботи операторів по збору боргів, 80% таких кредитів швидко повертається. Найбільш проблемна категорія – це шахраї. Шахрайство з споживчими кредитами зростає через те, що банки не обмінюються базами даних про позичальників.

Існують три основні способи роботи із “поганими” активами:

- Продаж простроченої заборгованості з балансу. Переваги: “Очищення” балансу банку; призупинення сплати податків на відсотки, що нараховуються; відсутність потреби працювати з простроченою заборгованістю. Недоліки: Необхідність вирішення питання про відповідальність за “сірі” схеми; неможливість дорого продати коротке прострочення.
- Аутсорсинг збору простроченої заборгованості. Переваги: Швидкий ефективний результат; можливість домовитися про гнучкі умови; можливість сконцентруватися на основній банківській діяльності; оплата лише за результат. Недоліки: Необхідність зробити “крок довіри”; відсутність стовідсоткового контролю за процесом збору заборгованості.
- Самостійний збір простроченої заборгованості. Переваги: Краще знання кредитного портфеля; досягнення ефективності за рахунок інтеграції збору заборгованості у внутрішні процеси банку; незалежність від зовнішніх колекторів. Недоліки: Необхідність істотних інвестицій. Тривалість процедури – щонайменше 12 місяців на організацію і налагодження процесів з моменту формування підрозділу по роботі з простроченою

заборгованістю; відволікання ресурсів банку на “боротьбу” з простроченою заборгованістю.

Продаж портфелів проблемної заборгованості є загальноприйнятою практикою у розвинених країнах, і останнім часом цей підхід набуває популярності і в Україні. Очевидно, що перед тим як щось продати чи купити, потрібно сформувати остаточну ціну товару, які задовольнила б обидві сторони. Але коли мова йде про портфель проблемної заборгованості, це є дуже нетривіальною задачею через вплив багатьох випадкових факторів. Саме тому важливим є процес розробки моделі оцінювання таких портфелів, яка б давала точні результати з урахуванням специфіки українського ринку.

1. ДОСЛІДЖЕННЯ КРЕДИТНОЇ СФЕРИ

1.1. Основні поняття

Перш за все необхідно чітко дати визначення усім економічним термінам, з якими ми будемо працювати.

Кредит – це кошти й матеріальні цінності, що надаються кредитором у користування позичальнику на визначений строк та під відсоток. Кредит поділяють на фінансовий, товарний і кредит під цінні папери, які засвідчують відносини позики.

Тип кредитного продукту – це певна класифікація кредитних продуктів в залежності від цілей кредитування, ризиків, що бере на себе банк, умов видачі кредиту та інших параметрів. Серед основних кредитних продуктів слід виділити наступні: грошовий кредит, кредит на споживчі потреби, заставні кредити (автокредит, іпотека), кредитні картки, мікрокредити, картки овердрафту. У даній роботі розглядаються портфелі, що складаються лише з перших двох типів кредитного продукту.

Кредитний портфель – це сукупність виданих банком кредитів, які на певну дату поки не погашені, тобто, знаходяться в користуванні позичальників. Видаючи кредити позичальникам, банк тим самим формує свій кредитний портфель.

Кредитний портфель банку становлять залишки коштів на балансових рахунках за короткостроковими, довгостроковими і простроченими кредитами, і цей показник відображає сукупність заборгованостей по активних кредитних операціях.

Колекторська агенція — спеціалізоване підприємство зі збору платежів (стягнення боргів).

Колекторські агенції переважно проводять стягнення на досудовому етапі існування заборгованості. Колектори забирають певний відсоток від

суми повернутого боргу залежно від його розміру та термінів заборгованості. Зазвичай колекторські агенції співпрацюють з кредитними установами (насамперед банками), а також з житлово-комунальними підприємствами, телекомунікаційними компаніями і навіть податковими органами.

Кількість днів прострочення платежу або DPD (Days Past Due) – це різниця у днях між датою виникнення останнього прострочення (день, у який мав бути черговий платіж) та поточною датою. У деяких випадках за DPD вважають різницю між датою виникнення прострочення та датою передачі договору у роботу колекторської компанії.

Дефолт першого платежу або FPD (First Payment Default) описує ситуацію, в якій прострочена заборгованість виникає одразу ж в рамках першого періоду користування кредитними коштами. Іншими словами, це відмова позичальника обслуговувати кредит або позику в перший місяць.

За даними фахівців, у багатьох випадках відмова погашати кредит з першого місяця пов’язана з шахрайством, тобто позичальник уже на етапі оформлення кредиту приймає рішення не повертати його. Дефолт першого платежу є великою проблемою для кредиторів, так як прострочення в перший місяць, як правило, говорить про те, що кошти повернути не вдасться. Велика частка FPD в портфелі банку свідчить про проблеми в скоринговій моделі компанії і вимагає термінового її перегляду і вдосконалення. В іншому випадку це може привести до банкрутства.

Сума виданого кредиту – це та сума, що була видана боржнику за умовами кредитного договору.

Компоненти боргу – це складові частини боргу, що передбачені кредитним договором. До них відносяться: тіло, проценти, комісія, штрафи, пеня та інші нарахування. Причому тіло, проценти і комісія бувають прострочені і непрострочені (залишки). Пеня і штрафи при цьому нараховуються лише на прострочені компоненти кредиту. Принципи

нарахування цих компонент є різними в залежності від банку і типу кредитного продукту.

Загальний розмір непогашеної заборгованості або сума до закриття – це сума всіх компонент боргу, які згідно кредитного договору боржник зобов'язаний повернути. Балансова вартість портфелю є рівною сумі усіх загальних розмірів непогашених заборгованостей усіх боржників, що входять до цього портфелю.

Загальна сума погашеної заборгованості – це показник, що демонструє скільки всього грошей сплатив боржник кредитору.

Термін кредиту – це період, на який видається кредит до його погашення. Короткостроковий, як правило, – на строк до одного року (переважно для формування оборотних коштів), середньостроковий – на строк до 5 років, і довгостроковий – понад 5 років, в основному як інвестиційний капітал.

Регіон фактичного проживання визначає ймовірне місце проживання боржника та дозволяє включити в оцінку певні макроекономічні показники (рівень безробіття, середня заробітна плата).

1.2. Визначення поняття проблемної заборгованості

Поняття проблемної заборгованості (або "поганого" активу) не є визначеним у законодавстві України, тому право ідентифікувати такі кредити та степінь їхньої "проблемності" передається у руки самих банків. У зв'язку з цим у фахівців існують багато різних але схожих визначень цього поняття [4], які можна представити у вигляді таблиці (табл. 1.1).

Таблиця 1.1 – Визначення поняття «проблемний кредит»

Автор	Визначення поняття «проблемний кредит»
М. Денисенко, В. Домрачев, В. Кабанов, В. Вовк, О. Хмеленко, У. Владичин	Кредит, за яким своєчасно не проведено один чи кілька платежів, значно знизилась ринкова вартість забезпечення, виникали обставини, за яких банк матиме сумнів щодо повернення позики [5, с. 105, 3, 4, 1]
О. Купчинова	Кредит, за яким установлено ознаки проблемності повернення, пов'язані з відсутністю або недостатністю забезпечення за кредитом, наявністю ознак фінансової нестійкості боржника або наявністю негативної інформації про його здатність виконати свої зобов'язання [7, с. 48]
Т. Осокина	Кредит, за яким існують серйозні потенційні та помірні реальні загрози, тобто мають місце утруднення у виконанні позичальником боргу [8, с. 5]
В. Кльоба	Кредит, за яким банк вбачає небезпеку своєчасного і повного його погашення внаслідок дії різноманітних чинників (економічних, юридичних, соціальних тощо) [9, с. 241]
Е. Шустова	Кредит, за яким позичальник не виконує зобов'язання (або виконує неналежним чином) у частині оплати платежів або є підстави вважати, що зобов'язання за ним не будуть виконані повністю або частково [10, с. 156]
С. Кузнецов	Кредит, за яким клієнт-боржник не здатний виконувати свої зобов'язання відповідно до прийнятих договорів та угод з банком, у зв'язку з чим існує потенційна загроза часткової або повної втрати для банку належних йому грошових коштів за кредитними зобов'язаннями боржника [11, с. 8]
О. Нурзат	Кредит, що має ряд ознак, з урахуванням яких він викликає у кредитних менеджерів обґрунтовані побоювання з приводу повернення основного боргу та відсотків за ним [12, с. 18]
Р. Хейнсворт, Е. Ніколаєнко, Л. Макаренко	Кредит, за яким позичальник вчасно не здійснив платіж або за яким існує висока ймовірність подібного неплатежу [13, с. 10]
О. Лаврушин	Кредит, за яким у банку виникли сумніви стосовно його суб'єкта, об'єкта та забезпечення [14, с. 381]
Н. Рабец	Кредит, за яким відсутні реальні джерела погашення, хоча строк погашення, можливо, ще не настав [15, с. 55]

Деякі з цих авторів ототожнюють проблемний кредит з простроченою заборгованістю, в той час як інші наголошують на низькій імовірності погашення кредиту з різних причин. Необхідно відзначити, що найбільш повними є ті визначення, в яких підсумовується можливість неповернення тіла кредиту та відсотків за його користування з будь-яких причин (зниження ринкової вартості забезпечення, погіршення фінансового стану позичальника тощо) або вже реальна ситуація прострочення за кредитом.

Отже, проблемним кредитом будемо вважати кредит, за яким своєчасно не проведено один чи кілька платежів, або виникли обставини, що викликають сумніви стосовно своєчасного та повного повернення наданого кредиту через фінансову нестійкість позичальника, недостатню забезпеченість чи незабезпеченість кредиту або з інших причин, що впливають на можливість неповернення кредиту та відсотків за його користування позичальником.

1.3. Огляд існуючих методів оцінювання кредитних портфелів

У цьому підрозділі будуть розглянуті два з багатьох існуючих методів оцінювання кредитних портфелів з проблемною заборгованістю.

1.3.1. Рамкова методика оцінки проблемних активів

Цей метод розроблений і описаний у [3]. Він являє собою оцінку кожного окремого кредиту з подальшою оцінкою портфеля в цілому. Результатом оцінки є розрахунок поверненої частини боргу. Колектору або

третій особі також доведеться врахувати час, необхідний для стягнення боргу, і використовувану колекторським агентством норму прибутковості капіталу.

По суті метод полягає в розгляді окремих кредитів в портфелі, а потім портфеля в цілому.

Завдання цієї методики полягає у визначенні властивостей, які:

- притаманні більшості кредитів;
- дозволяють прогнозувати можливу залишкову вартість кредиту.

Після чого кожній властивості присвоюється відповідне значення множника, помноживши який на номінальну вартість кредиту, ми отримаємо залишкову вартість окремого кредиту. Потім вартості окремих кредитів складаються і виходить вартість портфеля.

Схожий метод використовується і для портфеля кредитів, тобто визначаються характерні для портфеля властивості, і вартість портфеля множиться на відповідний множник. Якщо портфель до деякої міри однорідний, наприклад, всі кредити видані металообробним підприємствам, то множник, пов’язаний з даною галуззю економіки може бути «винесений за дужки» і визначений у якості множника для цілого портфеля. Іншими словами, у разі змішаних кредитів для кожного виду кредиту визначається окремий множник, але якщо портфель містить кредити однакового виду, то множник для всіх кредитів буде однаковим і буде портфельною властивістю.

Розглянемо безпосередньо формули, що використовуються у методі:

<Вартість фінансового активу> = *<Номінальна вартість>* * *<Значення фільтра>* * *<Індивідуальний множник>* (1.1)

*<Вартість застави> = менша з наступних величин: <Номінальна вартість> або <Номінальна вартість застави> * <Значення фільтра> * <Множник застави>*
(1.2)

<Вартість кредиту> = більша з наступних величин: <Вартість фінансового активу> або <Вартість застави>
(1.3)

<Вартість портфеля> = сума всіх <Вартість кредиту>, що входять в портфель
(1.4)

*<Справедлива вартість> = <Вартість портфеля> * <Портфельний множник> * <Показник якості кредитора>*
(1.5)

У цих формулах:

- Фінансовий актив – це вартість грошових потоків, яка підлягає поверненню позичальником.
- Номінальна вартість – це вартість, за якою проблемний кредит відображають на балансі банку без урахування резервів, виділених під нього банком, і без урахування застави.
- Застава – це інший актив, який можна продати для відшкодування зобов’язання позичальника перед кредитором.
- Номінальна вартість застави зазначається продавцем проблемного кредиту.
- Значення фільтра дорівнює 0 або 1, в залежності від виконання певних умов (описані нижче).

- Індивідуальний множник – це змінний множник, розрахований на основі знання конкретних умов окремого кредиту, а саме параметрів позичальника та фінансових умов кредиту.
- Заставний множник є аналогічним індивідуальному множнику, але для застави.
- Вартість кредиту – це залишкова вартість кредиту, а саме, або вартість фінансового активу, або вартість застави.
- Портфельний множник – це множник, що застосовується до всього портфелю. Може бути рівним індивідуальному множнику при однорідному портфелі кредитів.
- Показник якості кредитора – це показник, пов’язаний з кредитором; рейтинг кредитора.
- Справедлива вартість розраховується як номінальна вартість, помножена на індивідуальний і базовий множники з урахуванням можливого збитку.

У формулі (1.2) використана мінімальна функція, яка означає, що якщо вартість застави перевищить номінальну вартість кредиту, то при оцінці портфеля можна буде використовувати тільки вартість кредиту.

З формул можна зробити висновок, що продавець має обов’язково розкрити нам інформацію про:

- номінальну вартість кожного кредиту;
- номінальну вартість застави по кожному кредиту.

А наступні показники необхідно отримати використовуючи непряму інформацію про кредити і стан в економіці:

- значення фільтра;
- індивідуальні множники;
- заставний множник;

- портфельні множники;
- показник якості кредитора.

Цікавим є набір умов, який використовується автором методики для розрахунку значення фільтра. Згідно з цією методикою значення фільтра дорівнює 0, якщо відповідь на будь-яке з нижче вказаних питань є позитивною, і дорівнює 1, якщо відповіді на усі запитання негативні.

За фінансовим активом:

- Чи були випадки дефолту позичальника (у випадку з роздрібними клієнтами, чи має клієнт негативну кредитну історію)?
- Чи є факт несплати першого платежу по відсотках (це означає, що кредит був отриманий шахрайським способом)?
- Чи приймав банк заходи щодо кредитних фахівців, які видавали даний кредит?
- Чи є значна відмінність якості кредиту від якості інших кредитів з даного портфеля?
- Чи значно змінились дані про фінансовий стан позичальника при призначенні іншого кредитного фахівця для моніторингу кредиту?
- Чи відмовляється позичальник виплачувати кредит (це означає, що витрати по поверненню кредиту будуть високими внаслідок судових витрат)?

За заставою:

- Чи є комплект документації, що дозволяє стягнути заставу неповним?
- Чи є причина, що перешкоджає передачі прав на заставу від кредитора покупцеві портфеля?

За заставою і кредитом:

– Чи перевищує період часу між закінченням кредитної угоди і поточним днем тирічний термін, протягом якого не було подано судовий позов?

Також для прикладу наведемо пару таблиць з множниками для формул (табл. 1.2 і 1.3).

Таблиця 1.2 – Множники в залежності від виду кредиту

Вид кредиту	Множник (%)
Іпотека	70 - 75
Споживчий (виданий в точках продажу)	1 - 5
Автокредит (новий автомобіль або іномарка)	20 - 40
Автокредит (інший автомобіль)	1 - 5

Таблиця 1.3 – Множники в залежності від галузі економіки

Сектор економіки	Множник (%)
Видобувна промисловість	65 - 75
Металургія	65 - 75
Торгівля	25 - 35
Будівництво	15 - 25
Харчова промисловість	20 - 30
Лісне господарство	50 - 55
Фінансові компанії	10 - 15
Машинобудування	30 - 40
Інше	10 - 15

Автор цього методу також звертає увагу, що визначення і множники, включені в дану роботу, є орієнтовними і вимагають подальшої точної настройки перед застосуванням в комерційних цілях.

1.3.2. Оцінка за допомогою нейронних мереж

Нейронні мережі є дуже потужним засобом для розроблення методів для подібних задач оцінювання. Такий метод описано у статті одного з провідних спеціалістів у колекторській галузі [2].

Автор статті серед проблем відмічає проблему неповноти інформації про кредити, які оцінюються, і виділяє наступні параметри, як бажані при оцінці:

- EX_BODY – прострочена позичкова заборгованість ("тіло кредиту");
- EX_PERCENT – прострочені відсотки, комісії, штрафи та інші нарахування;
- CREDIT_AMOUNT – сума кредиту;
- CREDIT_MONTHS – термін, на який виданий кредит;
- BAD_DAYS – термін прострочення в днях (може бути замінений на дату виходу на прострочення);
- NICE_DAYS – кількість днів від видачі кредиту до виходу на прострочення (цей параметр найчастіше розраховується опосередковано, наприклад, знаючи дату видачі кредиту та дату виходу на прострочення);

– NICE_PAYMENTS – загальна сума платежів по кредиту, отриманих від позичальника (цей параметр відноситься до категорії "розкоші", так як його далеко не завжди включають до реєстрів для оцінки боргових портфелів, в той час як його інформативність виключно висока).

Найбільш принциповою вимогою щодо обміну інформацією при оцінці боргових портфелів є наявність в реєстрі полів EX_BODY і EX_PERCENT. Саме співвідношення цих двох полів дозволяє фахівцям (і комп'ютерним системам, в т.ч. скоринговим) виявити перспективність боргу для стягнення та оцінити можливу суму погашення. Ось деякі висновки, які можна зробити з аналізу цих двох величин:

1. За інших рівних умов, при покупці боргу слід оцінювати лише величину EX_BODY в якості бази для розрахунку вартості портфеля (статистика позасудового і судового стягнення всіляких штрафних санкцій та процентних нарахувань вкрай негативна).

2. Занадто велике відношення EX_PERCENT / EX_BODY (наприклад, більше 50%) означає, що боржнику нараховані підвищені відсотки і всілякі штрафні санкції. Це може свідчити про великий термін прострочення і свідоме ухилення боржника від погашення заборгованості, або про ситуацію, коли боржник не цілком розуміючи умов кредитного договору, допустив накопичення штрафних санкцій регулярно оплачуючи заборгованість недостатніми сумами. Практика показує, що таких "сумлінно-наївних" боржників буває надзвичайно важко переконати в необхідності погашення заборгованості за відсотками.

3. Відношення EX_PERCENT / EX_BODY близьке до нуля може свідчити про те, що незадовго перед продажем заборгованості банк отримав платежі від клієнта і направив їх (відповідно до пункту про пріоритетність зарахування платежів кредитного договору) на погашення пені, комісій і

відсотків. У таких випадках переконати клієнта про необхідність нових погашень на адресу нового власника заборгованості досить важко. З іншого боку, це може свідчити про контактність боржника, тобто про можливість проведення переговорів без необхідності проводити розшук.

Для заборгованостей з досить великим терміном прострочення (наприклад, BAD_DAYS > 200) важливу роль відіграє співвідношення між сумою платежів NICE_PAYMENTS і розміром кредиту CREDIT_AMOUNT. Можна вказати на дві закономірності:

1. Якщо NICE_PAYMENTS > CREDIT_AMOUNT, то психологічно буде дуже важко переконати боржника погасити залишок заборгованості, так як на його думку він давно виплатив суму кредиту.

2. Якщо NICE_PAYMENTS < 10% CREDIT_AMOUNT (звичайно, це нечітке співвідношення), то при значних термінах прострочення ми маємо явні ознаки несумлінного позичальника, який брав кредит без намірів погасити його.

На жаль, величина NICE_PAYMENTS не завжди доступна в реєстрах, пропонованих для оцінки. Тому автор пропонує використовувати перераховані вище параметри для непрямої оцінки сумлінності позичальника. На практиці можна з упевненістю сказати, що для заборгованостей з досить великим терміном прострочення можна стверджувати, що позичальник, який зробив менш одного - двох щомісячних платежів, є недобросовісним або неплатоспроможним. Оцінити кількість щомісячних платежів можна, знаючи значення NICE_DAYS, просто розділивши його на 30. Наприклад, якщо NICE_DAYS = 95, то орієнтовна кількість щомісячних платежів до виходу на прострочення складає 3.

Якщо параметр NICE_DAYS недоступний, можна оцінити кількість зроблених щомісячних платежів за приблизною формулою:

CREDIT_MONTHS * (CREDIT_AMOUNT - EX_BODY) /
CREDIT_AMOUNT

беручи до уваги той факт, що платежі від клієнта частково йдуть на погашення позичкової заборгованості.

Таким чином, з пропонованого компактного набору параметрів прострочення можна зробити ряд висновків про перспективи стягнення по кожному позичальнику і портфелю в цілому.

Крім об’єктивних даних про суми і терміни прострочення, можна спробувати ще докладніше проаналізувати представлений для оцінки реєстр простроченої заборгованості та виявити індивідуальні фактори ризику для окремих позичальників. Виходячи зі сформованої практики, а також принципу мінімалізму при обміні даними між контрагентами, автор пропонує використовувати для уніфікованого методу оцінки боргових портфелів наступні дані:

- DEBTOR_NAME – ПІБ боржника;
- CREDIT_DATE – дата видачі кредиту;
- CONTRACT_END – дата закінчення дії кредитного договору;
- LEGAL_ACTIONS – ознака подачі судового позову проти боржника.

Також корисними для оцінки є наступні дані:

- ADDRESS – адреса боржника;
- PHONES – телефони боржника;
- WORK – місце роботи боржника (найменування);
- WORK_PHONES – робочі телефони боржника.

Очевидно, що перелік додаткових відомостей може бути розширено.

Розглянемо деякі фактори ризику, які можна виявити з запропонованого набору параметрів:

1. Збіг DEBTOR_NAME в поточному реєстрі означає з високою ймовірністю, що один і той же позичальник допустив прострочення за кількома кредитними договорами в одному і тому ж банку. Очевидно, що це ознака несумлінності або неплатоспроможності даного позичальника за всіма кредитними договорами. Зауважимо також, що якщо в базі даних покупця боргового портфеля знаходяться прострочені борги тієї ж людини, це також свідчить про множинні боргові зобов'язання даної особи.

2. Збіг прізвищ боржників в одному і тому ж реєстрі може свідчити про можливу сімейну спорідненість боржників. Звичайно, збіг прізвищ, особливо поширених, не є 100% достовірною ознакою спорідненості, але слід привласнювати однофамільцям деяку підвищений ступінь ризику неповернення, тому що за статистикою множинні боргові зобов'язання в одній сім'ї свідчать про матеріальні труднощі, тобто про неплатоспроможність.

3. Збіг прізвищ і по батькові в одному і тому ж реєстрі з великою часткою ймовірності свідчить про родинні стосунки між боржниками (брат-сестра чи батько-син). За статистикою, множинні боргові зобов'язання по лінії братів-сестер мають високий ризик неповернення.

4. Виявлення одного з перерахованих вище факторів спільно з близькими або однаковими значеннями CREDIT_DATE свідчить про високий ризик неповернення. Виявлення цього фактора відповідає одній з наступних життєвих ситуацій: (1) отримання максимального числа кредитів в різних відділеннях банку однією і тією ж особою в один і той же день, з метою "обдурити" скорингову систему банку через повільний обмін даними при видачі кредитів різними відділеннями; (2) отримання членами однієї сім'ї або близькими родичами кредитів по одній з маркетингових програм банку в один і той же день: "Пішли, там гроші всім дають".

5. Приховані взаємозв’язки між позичальниками можна виявити, аналізуючи всю доступну інформацію перераховану вище. Наприклад, поширені шахрайські схеми (1) отримання множинних кредитів на співробітників одного підприємства (взаємозв’язок через найменування місця роботи або номера робочих телефонів). Яскравою ознакою шахрайства є збіг номера телефону, зазначеного одним боржником як "робочий телефон", а іншим боржником - в якості домашнього телефону.

Дані про юридичний статус кредитного договору слід використовувати для оцінки юридичних перспектив стягнення заборгованості. Так, якщо перевищений термін позовної давності за угодою (дата CONTRACT_END + 3 роки), і ніяких юридичних дій за угодою не зроблено (LEGAL_ACTIONS = 0), то перспективи як судового, так і позасудового стягнення мінімальні.

Таким чином, пропонований склад інформаційного реєстру для оцінки портфеля простроченої заборгованості дозволяє глибоко проаналізувати перспективи погашення і об’єктивно оцінити вартість портфеля.

Метою оцінки портфеля простроченої заборгованості є визначення суми погашення з урахуванням виявлених факторів ризику, а також оцінка витрат на управління портфелем і стягнення заборгованості. Різниця цих двох величин і дає об’єктивну величину вартості боргового портфеля, виражену в грошових одиницях.

Сума погашення прогнозується з урахуванням часового фактора, тобто має сенс розглядати функцію F = F(t).

З урахуванням багатофакторності критеріїв неплатоспроможності або кредитного шахрайства та загальної недетермінованності розглянутої задачі, для прогнозування суми погашення може використовуватися апроксимація функції F = F(t) за допомогою поліномів, нелінійних функцій або нейронних мереж. В даному випадку автором було обрано метод нейронних мереж.

В цьому випадку для прогнозування суми погашення по оцінюваному портфелю будемо подавати на вхід навченої нейронної мережі набір

параметрів, описаних вище, по черзі для кожного позичальника і на виході отримувати індивідуальну оцінку суми погашення. Загальна сума таких індивідуальних прогнозів дасть значення суми погашення по всьому оцінюваному портфелю.

Важлива проблема при побудові нейронних мереж і будь-яких інших апроксимуючих функцій - попередня обробка даних. Інтуїтивно зрозуміло, що не завжди абсолютне значення суми заборгованості має достатню прогностичну силу: і справді, якщо людина вийшла на прострочення 180 днів тому, чи так уже важливо, винен він банку 400 або 500 тис. гривень? Імовірність погашення не буде прямо залежати від точного значення: головне, що людина винна "велику суму грошей". А ось якщо при такому ж терміні прострочення сума заборгованості становить 10 тис. гривень, ситуація кардинально змінюється. Імовірність стягнення "малої суми" значно зростає.

Для математичного опису якісних показників, таких як "велика заборгованість" і "мала заборгованість" використовується апарат нечітких множин. Наприклад, на рис. 1.1 зображені функції приналежності двох нечітких множин "мала" і "велика" заборгованість, які словесно можна описати як "заборгованості близькі до нуля" і "заборгованості близькі до 500 тис. гривень." По осі абсцис - значення заборгованості в гривнях, по осі ординат - ступінь приналежності до першої або другої нечіткої множини (вимірюється від 0 до 1).

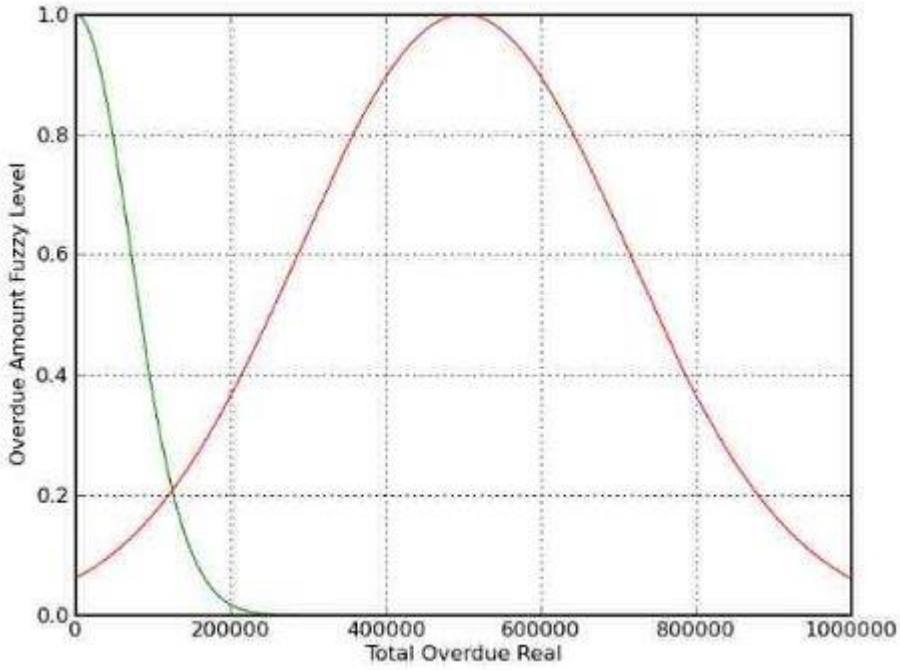

Рисунок 1.1 – Функції належності для нечітких множин “мала сума” (зеленим) і “велика сума” (червоним)

На підставі дослідження накопичених статистичних даних, використання ступенів належності до нечітких множин в якості вхідних параметрів для нейронної мережі значно покращує якість прогнозу.

Підсумовуючи, можна описати запропоновану методику оцінки по кроках:

1. Використовувати для оцінки набір параметрів, описаний вище.
2. Перетворити параметри в ступені належності до нечітких множин, таких як "велика заборгованість" або "малий термін прострочення" або "середній степінь взаємозв’язку суб’єктів".
3. Побудувати нейронну мережу, яка видає прогноз погашення по кожному позичальнику на підставі параметрів, підготовлених в п. 2, у вигляді функції часу F = F(t).
4. Розрахувати прогноз погашення по кожному позичальнику з оцінюваного портфеля.

5. Сума прогнозів, отриманих в (4), дає підсумковий прогноз погашення по оцінюваному портфелю.

1.3. Висновку до розділу 1

В цілому, можна зробити висновок, що задача оцінювання портфелів проблемної заборгованості є актуальною в Україні, і, оскільки колекторська галузь в Україні все ще тільки розвивається, методи, які наразі використовуються, не є ідеальними.

Було розглянуто два таких методи. Перший – зовсім простий і заснований на статистичних показниках, виведених емпіричним шляхом, а другий, що використовує нейронні мережі, – більш розвинений і може використовуватись як робочий інструмент.

2. РІЗНОВИДИ ЛІНІЙНОЇ РЕГРЕСІЇ

2.1. Класична регресія

Для розв’язання задачі поставленої у цій роботі будуть використовуватися методи з розділу регресійного аналізу, тому буде доречним розглянути основи цієї теорії.

2.1.1. Означення регресії

Для початку потрібно зрозуміти, що таке регресія. Нехай Y та X_1, X_2, ..., X_p – це деякі випадкові величини; y, x_1, x_2, … , x_p – це фіксовані значення відповідних випадкових величин; а y^i, x_1^i, x_2^i, … , x_p^i $(i = 1 \dots n)$ – це конкретні значення кожної з випадкових величин для i-го спостереження.

Можна зробити припущення, що величини X_j $(j = 1 \dots p)$ впливають на значення величини Y і записати таку залежність у вигляді рівняння:

$$f(x_1, \dots, x_p) = E(Y / X_1 = x_1, \dots, X_p = x_p). \quad (2.1)$$

Воно вводить нову функцію f, яка визначає математичне сподівання Y при умові, що величини X_j $(j = 1 \dots p)$ – фіксовані і дорівнюють певним значенням x_j $(j = 1 \dots p)$. При цьому Y називають залежною (або критеріальною) величиною, а X_j $(j = 1 \dots p)$ – незалежними величинами (або регресорами чи предикторами). Рівняння (2.1) називають рівнянням регресії в загальному вигляді, а саму функцію f називають **регресією**

величини Y по величинам X_j $(j = 1 \ldots p)$. Якщо функція f лінійно залежить від своїх параметрів, то регресію також називають лінійною.

Для прикладу розглянемо залежність ваги від зросту серед чоловіків. Нехай були досліджені три пацієнти: один важить 81 кг при зрості 191 см, другий – 75 кг при 180 см, і третій – 93 кг при 200 см. Будемо вважати вагу – залежною величиною Y, а зріст – незалежною величиною X (в даному випадку незалежна величина лише одна). Тоді спостереження можна записати у вигляді векторів:

$$Y = \begin{pmatrix} y^1 \\ y^2 \\ y^3 \end{pmatrix} = \begin{pmatrix} 81 \\ 75 \\ 93 \end{pmatrix}, \; X = \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} 191 \\ 180 \\ 200 \end{pmatrix}. \tag{2.2}$$

Якщо припустити, що залежність між цими величинами лінійна, то функцію f можна записати у вигляді:

$$f(x) = \beta_0 + \beta_1 x, \tag{2.3}$$

де β_0 – це вільний коефіцієнт (математичне сподівання значення величини Y при нульовому значенні незалежної змінної), а β_1 – це лінійний коефіцієнт при незалежній змінній (або наскільки змінюється значення залежної змінної при збільшенні цієї незалежної змінної на одиницю).

2.1.2. Метод найменших квадратів (МНК)

Виходячи з рівняння регресії (2.1) можна записати наступне:

$$Y = f(X) + E, \tag{2.4}$$

де E – це випадкова величина з нульовим математичним сподіванням, що відіграє роль похибки.

У більшості випадків ми не знаємо справжній вигляд функції $f(X)$, але ми можемо припустити, що вона є лінійною:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2.5}$$

Позначимо:

$$\hat{Y} \overset{\text{def}}{=} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = X\beta. \tag{2.6}$$

$\hat{Y}$ – це лінійна апроксимація залежної змінної, розрахована на основі незалежних змінних.

На основі попередніх трьох формул можна записати:

$$E = Y - \hat{Y}, \tag{2.7}$$

$$E^T E = (Y - \hat{Y})^T (Y - \hat{Y}). \tag{2.8}$$

Якщо перейти до фіксованих значень випадкових величин, то рівняння (2.8) перетворюється на:

$$\sum_{i=1}^{n} \left(e^i\right)^2 = \sum_{i=1}^{n} \left(y^i - \hat{y}^i\right)^2 = \sum_{i=1}^{n} \left(y^i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j^i \right)^2. \tag{2.9}$$

Метод найменших квадратів (МНК) у застосуванні до лінійної регресії полягає у пошуку таких коефіцієнтів β_j $(j = 1 \dots p)$, які б мінімізували похибку $\sum_{i=1}^{n}(e^i)^2$:

$$(Y - X\beta)^T(Y - X\beta) \to \min_\beta \ . \tag{2.10}$$

Неважко побачити, що рішення цієї задачі зводиться до рішення системи рівнянь:

$$X^T X\beta = X^T y \ \Rightarrow \ \beta = (X^T X)^{-1}X^T y. \tag{2.11}$$

При цьому для останнього перетворення матриця $X^T X$, очевидно, має бути невиродженою.

Продовжимо приклад про залежність ваги від зросту. Виходячи з (2.2) отримуємо:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 \\ 191 & 180 & 200 \end{pmatrix}\begin{pmatrix} 1 & 191 \\ 1 & 180 \\ 1 & 200 \end{pmatrix} = \begin{pmatrix} 3 & 571 \\ 571 & 108881 \end{pmatrix}. \tag{2.12}$$

Обернена до неї матриця має вигляд $(X^T X)^{-1} = \frac{1}{602}\begin{pmatrix} 108881 & -571 \\ -571 & 3 \end{pmatrix}$.

Враховуючи (2.11) знаходимо:

$$\beta = \frac{1}{602}\begin{pmatrix} 108881 & -571 \\ -571 & 3 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 191 & 180 & 200 \end{pmatrix}\begin{pmatrix} 81 \\ 75 \\ 93 \end{pmatrix} \approx \begin{pmatrix} -85{,}834 \\ 0{,}887 \end{pmatrix}. \tag{2.13}$$

Таким чином ми отримали апроксимацію $\hat{y} = -85{,}834 + 0{,}887x$. Результати свідчать про те, що залежність між вагою та зростом є прямо

пропорційною, що й не дивно, і про те, що на кожен сантиметр зросту вага чоловіка збільшується в середньому на 0,887 кг.

2.1.3. Статистичні властивості МНК-оцінок

Для того, щоб лінійна оцінка, отримана за допомогою МНК, була незміщеною достатньо виконання таких умов:

- Математичне сподівання похибки E має дорівнювати нулю. Цим пунктом можна знехтувати, якщо в оцінку включений вільний член, оскільки він «візьме на себе» зміщення оцінки.
- Похибка E має бути незалежно розподіленою випадковою величиною від предикторів X_j $(j = 1 \dots p)$.

Якщо ми хочемо, щоб оцінка була ще й ефективною (тобто найкращою в класі незміщених оцінок), необхідні виконання таких обмежень на розподіл похибки:

- Дисперсія E має бути постійною для всіх спостережень (тобто має виконуватись умова рівномірності дисперсії – так званої гомоскедастичності): $D(\varepsilon^i) = \sigma^2 = const.$
- Має бути відсутня кореляція випадкових похибок в різних спостереженнях між собою: $cov(\varepsilon^i, \varepsilon^j) = 0,\ i,j = 1 \dots n,\ i \neq j.$

Лінійна модель, що задовольняє всі ці умови називаються класичною. Оцінка отримана за допомогою такою моделі є незміщеною, конзистентною і ефективною.

Окрім цього важливо також, щоб значення залежної змінної були нормально розподілені при фіксованих значеннях незалежних змінних.

2.2. Узагальнена лінійна регресія

У попередньому розділі були розглянуті лінійні моделі, які можна використовувати для передбачення значень залежних змінних з нормальним розподілом по значенням набору неперервних і/або категоріальних незалежних змінних. Однак часто у нас немає ніяких підстав припускати, що залежні змінні нормально розподілені (і навіть неперервні). Наприклад:

- Залежна змінна може бути категоріальною. Бінарні змінні (наприклад, так / ні, здав / провалився, живий / мертвий) і категоріальні змінні (наприклад, поганий / хороший / чудовий, республіканець / демократ / незалежний) точно не характеризуються нормальним розподілом.
- Залежна змінна може бути лічильною (наприклад, число дорожньо-транспортних пригод за тиждень, число порцій спиртного в день). Такі змінні мають обмежене число значень і ніколи не бувають негативними. Крім того, їх математичне сподівання і дисперсія часто пов’язані (це не виконується для змінних з нормальним розподілом).

Узагальнені лінійні моделі розширюють застосовність лінійних моделей, роблячи можливим аналіз залежних змінних, що мають відмінний від нормального розподіл.

2.2.1. Сімейство експоненційних розподілів

Узагальнена лінійна модель, являє собою цілий комплекс моделей, серед яких класична модель міститься як окремий випадок. При цьому передбачається, що компоненти спостережуваного вектора можуть мати різні імовірнісні розподіли, які належать сімейству експоненційних розподілів. До цього сімейства належать нормальний розподіл, розподіл Пуассона, гамма розподіл, біноміальний розподіл та інші. Крім того, у компонент спостережуваного вектора може бути різна дисперсія.

Сімейство експоненційних розподілів з одним скалярним параметром – це усі розподілі, функція щільності (або функція ймовірності для дискретного випадку) яких може бути представлена у вигляді:

$$f_X(x|\theta) = h(x)\exp(\eta(\theta)T(x) - A(\theta)), \tag{2.14}$$

де $T(x)$, $h(x)$, $\eta(\theta)$ та $A(\theta)$ – це відомі функції.

Також розповсюджена альтернативна форма цієї формули:

$$f_X(x|\theta) = h(x)g(\theta)\exp(\eta(\theta)T(x)). \tag{2.15}$$

Значення θ називають параметром сімейства.

Крім того, на носій (support) функції $f_X(x|\theta)$ (тобто множина таких значень x, де $f_X(x|\theta)$ більша за нуль) не може залежати від параметра θ. Це обмеження часто використовують для відсікання розподілів, що не належать до експоненційного сімейства. Наприклад розподіл Парето не належить до нього, оскільки він ненульовий тільки на множині $x > x_m$, де x_m – це параметр розподілу.

Слід також зазначити, що функція $A(\theta)$ або її еквівалент $g(\theta)$ автоматично визначені, якщо визначені всі інші функції, оскільки для всіх функцій щільності має виконуватись умова нормування (інтеграл по всій області визначення має дорівнювати одиниці). При цьому можна показати, що функцію $A(\theta)$ завжди можна виразити через η (2.19).

Однак сімейство експоненційних розподілів було в дуже бідним, якби допускався лише скалярний параметр θ, тому допускається параметр-вектор $\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_s \end{pmatrix}$. При цьому, зрозуміло, що функція $\eta(\theta)$ також має стати вектор-функцією $\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \eta_1(\boldsymbol{\theta}) \\ \vdots \\ \eta_s(\boldsymbol{\theta}) \end{pmatrix}$, а формула (2.14) перетвориться на:

$$f_X(x|\boldsymbol{\theta}) = h(x)\exp\left(\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta})T_i(x) - A(\boldsymbol{\theta})\right). \tag{2.16}$$

Це можна записати більш компактно:

$$f_X(x|\boldsymbol{\theta}) = h(x)\exp(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{T}(x) - A(\boldsymbol{\theta})). \tag{2.17}$$

Формула (2.15) перетвориться на:

$$f_X(x|\boldsymbol{\theta}) = h(x)g(\boldsymbol{\theta})\exp\big(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{T}(x)\big). \tag{2.18}$$

Зазвичай розмірності векторів $\boldsymbol{\theta}$ і $\boldsymbol{\eta}(\boldsymbol{\theta})$ співпадають, але деякі розподіли характеризуються параметром $\boldsymbol{\theta}$, що має меншу розмірність, ніж $\boldsymbol{\eta}(\boldsymbol{\theta})$. Такі члени сімейства експоненційних розподілів називаються «викривленими» (curved).

Зазначимо також, що $\boldsymbol{\eta}(\boldsymbol{\theta})$ називають «натуральним параметром» (natural parameter), а $\boldsymbol{T}(x)$ – «достатньою статистикою» (sufficient statistic).

Функцію $A(\boldsymbol{\eta})$ називають лог-нормальним фактором (log-partition function), тому що це натуральний логарифм нормалізуючого фактора, без якого $f_X(x|\boldsymbol{\theta})$ не може бути щільністю ймовірнісного розподілу:

$$A(\boldsymbol{\eta}) = \ln\left(\int_x h(x) exp\big(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \boldsymbol{T}(x)\big)\, dx\right). \tag{2.19}$$

2.2.2. Представники експоненційного сімейства

Дуже важливо зазначити, що ключовим фактором, який впливає на те, чи є розподіл експоненційним, є його параметри, а саме – які з цих параметрів є фіксованими, а які можна вважати змінними (частиною вектора $\boldsymbol{\theta}$). Так, наприклад, біноміальний розподіл не є експоненційним, якщо його параметр n (кількість випробувань) є змінним, оскільки це суперечить обмеженню на носій функції $f_X(x|\theta)$. Однак трохи далі буде показано що, якщо змінним вважати лише параметр p (імовірність успіху), то біноміальний розподіл належить до сімейства експоненційних.

Для початку розглянемо нормальний розподіл із фіксованою дисперсію σ^2 та змінним математичним сподіванням μ. Щільність розподілу виглядає наступним чином:

$$f_\sigma(x;\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.20}$$

Легко показати, що вона має форму (2.14), якщо взяти:

$$h_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$

$$T_\sigma(x) = \frac{x}{\sigma},$$

$$A_\sigma(\mu) = \frac{\mu^2}{2\sigma^2} \text{ або } A_\sigma(\eta) = \frac{\eta^2}{2},$$

$$\eta_\sigma(\mu) = \frac{\mu}{\sigma}.$$

Перейдемо тепер до випадку, коли і дисперсія σ^2, і математичне сподіванням μ є змінними. Щільність має вигляд:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2.21}$$

а характеризуючі функції підібрані наступним чином:

$$h(x) = \frac{1}{\sqrt{2\pi}},$$

$$\boldsymbol{T}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix},$$

$$A(\mu,\sigma) = \frac{\mu^2}{2\sigma^2} + \ln|\sigma| \text{ або } A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\ln\left|\frac{1}{2\eta_2}\right|,$$

$$\boldsymbol{\eta}(\mu,\sigma) = \begin{pmatrix} \dfrac{\mu}{\sigma^2} \\ -\dfrac{1}{2\sigma^2} \end{pmatrix}.$$

І нарешті, розглянемо біноміальний розподіл з фіксованим числом випробувань n та змінною ймовірністю успіху p. Він є дискретним, тому треба використати функцію імовірності:

$$f(x) = C_n^x p^x (1-p)^{n-x}, x \in (1,2,\dots,n). \quad (2.22)$$

Це є тим самим, що й:

$$f(x) = C_n^x \exp\left(x\, ln\left(\frac{p}{1-p}\right) + n\, ln(1-p)\right), x \in (1,2,\dots,n). \quad (2.23)$$

Отже, біноміальний розподіл також належить до сімейства експоненційних з натуральним параметром $\eta(p) = ln\left(\frac{p}{1-p}\right)$. Ця функція відома від назвою «логіта» (logit) і є оберненою до логістичної функції:

$$\text{logit}^{-1}(p) = \text{logistic}(p) = \frac{1}{1+e^{-p}} = \frac{e^p}{e^p+1}.$$

2.2.3. Метод максимальної правдоподібності

Метод максимальної правдоподібності у математичній статистиці – це метод оцінювання невідомого параметра шляхом максимізації функції правдоподібності. Він ґрунтується на припущенні про те, що вся інформація про статистичну вибірку міститься у цій функції.

Функція правдоподібності, у свою чергу, – це спільний розподіл вибірки з параметричного розподілу, що розглядається як функція від параметра. При цьому використовується спільна функція щільності у випадку неперервного розподілу і спільна функція ймовірність у випадку дискретного розподілу.

Найкраще можна зрозуміти цей метод на прикладі. Припустимо, ми спостерігаємо за експериментом з підкидання монети з ймовірністю

випадіння аверсу p, що суттєво відрізняється від 50%. Необхідно на основі послідовності випадінь аверсів і реверсів, оцінити істинне значення p.

Для початку необхідно знайти спільну функцію ймовірності для цієї серії випадкових подій. Результат кожного підкидання монети можна вважати розподіленим за законом Бернуллі. Функція ймовірностей цього розподілу виглядає наступним чином:

$$P(x;p) = p^x(1-p)^{1-x}, \tag{2.24}$$

де $x = 1$, якщо випав аверс, і $x = 0$, якщо випав реверс.

Функція правдоподібності для серії зі ста експериментів, результати яких розподілені за законом Бернуллі, в такому випадку буде просто результатом добутку усіх функцій ймовірності:

$$L(p) = \prod_{i=1}^{100} p^{x_i}(1-p)^{1-x_i}, \tag{2.25}$$

де x_i – це результат i-го випробування.

Нехай після ста кидків аверс випав 29 разів, а реверс – 71. Знаючи це, ми тепер можемо використати метод максимальної правдоподібності для того, щоб, наприклад, визначити, яке зі значень p найбільш доречне по відношенню до результатів випробувань: $p = \frac{1}{4}$, $p = \frac{1}{3}$ або $p = \frac{1}{2}$.

Для цього підрахуємо значення функції правдоподібності для різних значень p:

$$L(p) = \prod_{i=1}^{100} p^{x_i}(1-p)^{1-x_i} = p^{29}(1-p)^{71},$$

$$L\left(\frac{1}{4}\right) = \left(\frac{1}{4}\right)^{29}\left(\frac{3}{4}\right)^{71} \approx 4{,}67 \cdot 10^{-24},$$

$$L\left(\frac{1}{3}\right) = \left(\frac{1}{3}\right)^{29}\left(\frac{2}{3}\right)^{71} \approx 4{,}58 \cdot 10^{-27},$$

$$L\left(\frac{1}{4}\right) = \left(\frac{1}{2}\right)^{29}\left(\frac{1}{2}\right)^{71} \approx 7{,}89 \cdot 10^{-31}.$$

Видно, що функція досягає максимального значення при $p = \frac{1}{4}$, що і є відповіддю на поставлене запитання.

Але що, якщо нам необхідно знайти найкраще значення p серед усіх можливих? В такому разі, очевидно, треба розв'язати задачу оптимізації:

$$\max_{0 \le p \le 1} L(p) = \max_{0 \le p \le 1} \prod_{i=1}^{100} p^{x_i}(1-p)^{1-x_i} = \max_{0 \le p \le 1} p^{29}(1-p)^{71}.$$

Продифиренціюємо функцію по параметру p:

$$L'(p) = 29p^{28}(1-p)^{71} - 71p^{29}(1-p)^{70} =$$
$$= p^{28}(1-p)^{70}(29(1-p) - 71p) = p^{28}(1-p)^{70}(29 - 100p).$$

Після чого прирівняємо цей вираз до нуля і розв'язавши рівняння отримаємо можливі значення для p: 0, 1 та 0,29. Значення 0 та 1 не підходять, оскільки вони перетворюють функцію максимальної правдоподібності на нуль. Отже серед усіх можливих значень параметра, найкращим є $p = \frac{29}{100}$, що є доволі передбачуваним результатом.

Далі розглянемо приклад з неперервним розподілом. Нехай ми спостерігаємо за серією з n незалежних випробувань, результат кожного з яких розподілений нормально з математичним сподіванням μ і дисперсією σ^2.

Як і в попередньому випадку, ми хочемо знайти оцінку для параметрів розподілу. Для того, щоб спростити задачу, будемо вважати, що дисперсія σ^2 є відомою, а нам треба оцінити лише μ.

Щільність нормального розподілу записується як:

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.26}$$

Спільна функція щільності, що і є функцією правдоподібності, для серії з n незалежних однаково розподілених нормальних величин є добутком щільностей:

$$L(\mu) = f(x_1, \dots, x_n; \mu) = \prod_{i=1}^{n} f(x_i;\mu) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right).$$

Якщо ввести позначення $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, то можна переписати це наступним чином:

$$L(\mu) = f(x_1, \dots, x_n; \mu) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right).$$

Для того, щоб максимізувати цю функцію по μ, потрібно використати той факт, що логарифмічна функція є неперервною і монотонною строго зростаючою функцією на всій осі, тому максимізувати функцію правдоподібності – це те саме, що максимізувати її логарифм. Знайдемо його:

$$\ln\big(L(\mu)\big) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Тепер продифиренціюємо цей вираз по μ і отримаємо:

$$\frac{d}{d\mu}\ln\big(L(\mu)\big) = 0 - \frac{-2n(\bar{x}-\mu)}{2\sigma^2} = \frac{n}{\sigma^2}(\bar{x}-\mu).$$

Прирівнявши це до нуля ми отримаємо відповідь: $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Цікаво, що результат виявився незалежним від значення σ^2.

2.2.4. Використання методу УЛР

Як було зазначено, узагальнена лінійна регресія використовується, коли вихідна змінна розподілена не нормально. Для цього припускається, що математичне сподівання залежної змінної залежить від незалежних таким чином:

$$E(\boldsymbol{Y}) = \boldsymbol{\mu} = g^{-1}(\boldsymbol{X\beta}), \qquad (2.27)$$

де $E(\boldsymbol{Y})$ – це математичне сподівання $\boldsymbol{Y}$, $\boldsymbol{X\beta}$ – лінійна комбінація предикторів, а g – зв'язуюча функція.

Зв'язуюча функція надає зв'язок між лінійною комбінацією предикторів і середнім функції розподілу. Використовується багато різних функцій зв'язку. Вибір конкретної функції зв'язку залежить від пропонованого розподілу залежної змінної. У таблиці 2.1 представлені основні найчастіше використовувані функції.

Таблиця 2.1 – Приклади зв’язуючих функцій

Розподіл	Застосування	Назва зв’язку	Функція зв’язку	Обернена ф-ія
Нормальний	Лінійний відгук	Identity	$g(\mu) = \mu$	$g^{-1}(\boldsymbol{X\beta}) = \boldsymbol{X\beta}$
Експоненційний	Експоненційний відгук	Negative inverse	$g(\mu) = -\frac{1}{\mu}$	$g^{-1}(\boldsymbol{X\beta}) = = -\frac{1}{\boldsymbol{X\beta}}$
Пуассона	Кількість подій за проміжок часу	Log	$g(\mu) = \ln(\mu)$	$g^{-1}(\boldsymbol{X\beta}) = = \exp(\boldsymbol{X\beta})$
Бернуллі	Результат однієї «так/ні» події	Logit	$g(\mu) = = \ln\left(\frac{\mu}{1-\mu}\right)$	$g^{-1}(\boldsymbol{X\beta}) = = \frac{1}{1+\exp(\boldsymbol{X\beta})}$
Біноміальний	К-ть «так» подій серед n «так/ні» подій			

Отже, для того, щоб побудувати, наприклад, модель для передбачення кількості певних подій за фіксований проміжок часу, необхідно підігнати модель:

$$\ln(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \,,$$

де μ – це середня кількість подій за заданий проміжок часу.

Оцінюються такі системи за допомогою методу максимальної правдоподібності, який в свою чергу апроксимується чисельними методами, такими як ітеративний МНК зі змінними вагами та метод Ньютона-Рафсона.

2.3. Логістична регресія та ROC-аналіз

Логістична регресія є частковим випадком узагальненої лінійної регресії. Вона використовується у задачах, де вихідна змінна є бінарною або категоріальною, бо в таких випадках використання класичної лінійної регресії недопустиме. Логістична регресія є одним з методів бінарної класифікації. Вона дозволяє оцінити ймовірність настання (або не настання) події в залежності від значень незалежних змінних.

2.3.1. Опис методу логістичної регресії

Ми розглядаємо модель з декількома незалежними змінними $X_1, \dots, X_p$ і однією залежною бінарною змінною Y, що може приймати лише значення 1 («так») та 0 («ні»). В такому випадку можна сказати, що вихідна величина розподілена за законом Бернуллі.

Нехай $E(Y) = \mu$ – це математичне сподівання, тобто ймовірність того, що залежна змінна прийме значення 1. В якості функції зв'язку оберемо логіту. Тоді модель виглядатиме наступним чином:

$$\ln\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \ . \qquad (2.28)$$

Після підгонки моделі, знайшовши всі коефіцієнти $\beta_0, \dots, \beta_p$, для кожного набору незалежних змінних ми можемо апроксимувати значення $\ln\left(\frac{\mu}{1-\mu}\right)$.

Для того, щоб краще розуміти сенс значень, отриманих за допомогою цієї моделі, перепишемо рівняння (2.28) у наступному вигляді:

$$\frac{\mu}{1-\mu} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} = e^{\beta_0} e^{\beta_1 x_1} \dots e^{\beta_p x_p} . \qquad (2.29)$$

Як вже було сказано μ – це ймовірність того, залежна змінна Y прийме значення 1 («так» або «успіх»). Таким чином, у лівій частині (2.29) ми маємо так звані «шанси» («odds»), тобто відношення можливості успіху до можливості невдачі. Наприклад, шанси $1:1$ $(1=1:1)$ означають 50% імовірність успіху, а $4:1$ $(0{,}25=4:1)$ – 80%.

У правій частині (2.29) ми бачимо, що кожна незалежна змінна окремо впливає на залежну, а саме вона помножує шанси у певну кількість разів. Наприклад, якщо β_1 дорівнює 2, то кожне збільшення x_1 на одиницю збільшує шанси у $e^2 \approx 7{,}39$ разів, а якщо β_2 дорівнює -3, це означає зменшення шансів у $e^3 \approx 20{,}09$ разів на кожне збільшення x_2 на одиницю.

2.3.2. Основи ROC-аналізу

У задачах бінарної класифікації, коли модель передбачає ймовірність того, що результат спостереження відноситься до одного з двох класів, дуже важливий вибір точки відсікання, тобто порогу ймовірності, що розділяє два класи. Така точка відсікання показує, після якого значення ймовірності на виході моделі один клас змінюється іншим. Вибираючи точку відсікання, ми управляємо ймовірністю правильного розпізнавання позитивних і негативних прикладів. При зменшенні порогу відсікання збільшується ймовірність помилкового розпізнавання позитивних спостережень (хибнопозитивних

результатів), а при збільшенні зростає ймовірність неправильного розпізнавання негативних спостережень (хибнонегативних результатів).

Мета ROC-аналізу полягає в тому, щоб підібрати таке значення точки відсікання, яке дозволить моделі з найбільшою точністю розпізнавати позитивні або негативні приклади і видавати найменшу кількість хибнопозитивних або хибнонегативних помилок, відповідно.

Опишемо сказане більш формально. Нехай ми спостерігаємо за серією з n експериментів, результати яких можуть бути позитивним (P) або негативними (N). Припустимо кожен з експериментів має результат $X_i, i = 1 \dots n$, тобто кожен з X_i дорівнює або P, або N. Нехай також ми маємо модель, що може передбачати ймовірність результату кожного експерименту. Для кожного випробування модель видає число $p_i \in [0,1], i = 1 \dots n$, що є ймовірністю того, що i-те випробування має позитивний результат P.

Для того, щоб інтерпретувати результати моделі, необхідно ввести значення порогу $T \in [0,1]$. Так, якщо $p_i \leq T$, то передбаченням для i-го випробування будемо вважати N, а якщо $p_i > T$, – то P. Позначимо кожен такий результат як $\hat{X}_i, i = 1 \dots n$.

Таким чином для кожного експерименту ми маємо передбачення $\hat{X}_i$ і справжній результат X_i. Можливі чотири комбінації значень X_i і $\hat{X}_i$: PP, PN, NP і NN. PP відповідає ситуації, коли передбачення є правильним і правдивим, а NN – коли передбачення є правильним і негативним. У випадку NP ми маємо неправильне передбачення, оскільки модель передбачила позитивний результат, в той час як він є негативним. В такому разі кажуть, що була зроблена помилка першого роду. Відповідно, у разі PN ми отримуємо помилку другого роду.

Такі результати прийнято відображати у вигляді таблиці, яку називають таблицею спряженості або таблицею помилок (табл. 2.2).

Таблиця 2.2 – Таблиця помилок

Усього спостережень (n)		Істина	
		Справжнє значення позитивне	Справжнє значення негативне
Передбачення	Передбачення позитивне	PP (true positive (TP))	NP (false positive (FP) або помилка 1-го роду)
	Передбачення негативне	PN (false negative (FN) або помилка 2-го роду)	NN (true negative (TN))

На основі цих кількісних показників можна розрахувати багато інформативних характеристик моделі, які використовуються у ROC-аналізі. Основними з них є чутливість (Se) і специфічність (Sp).

Чутливість (Se), також відома як TPR (true positive rate), – це частка правильно передбачених позитивних результатів серед усіх позитивних випадків:

$$Se = TPR = \frac{TP}{TP + FN}\,. \tag{2.30}$$

Більша чутливість моделі означає більшу можливість передбачення бути позитивним, тобто меншу можливість зробити помилку 2-го роду.

У свою сергу, специфічність (Sp), також відома як TNR (true negative rate) або $1 - FPR$ (false positive rate), – це частка правильно передбачених негативних результатів серед усіх негативних випадків:

$$Sp = TNR = 1 - FPR = \frac{TN}{TN + FP}\,. \tag{2.30}$$

Більша специфічність моделі означає меншу можливість зробити помилку 1-го роду.

Це означає, що модель з високою чутливістю часто дає правильний результат для позитивних випадків (вона більше на них спеціалізується), а модель з високою специфічністю краще працює на негативних випадках. Для кращого розуміння цих концептів можна спробувати зрозуміти, що вони будуть означати в термінах проблеми діагностування медичних пацієнтів. Чутливий діагностичний тест призводить до гіпердіагностики – максимального запобігання пропуску хворих, і підвищує можливість помилки 1-го роду. Специфічний тест буде діагностувати лише достовірно хворих. Це може бути корисним, якщо побічні ефекти від лікування є занадто небезпечними для помилково діагностування.

2.3.3. ROC-крива

ROC-крива – це основний інструмент ROC-аналізу. Будується вона наступним чином:

1. Для кожного можливого значення порогу відсікання T від 0 до 1 з певним кроком dT (наприклад $dT = 0{,}01$) розраховуються значення чутливості (Se) і специфічності (Sp).
2. Будується графік, на осі Y якого відкладаються значення Se, а на осі X відкладаються $1 - Sp$, розраховані у попередньому пункті.

У результаті виходить деяка параметрична крива, що, як правило, виглядає як крива на рис. 2.1.

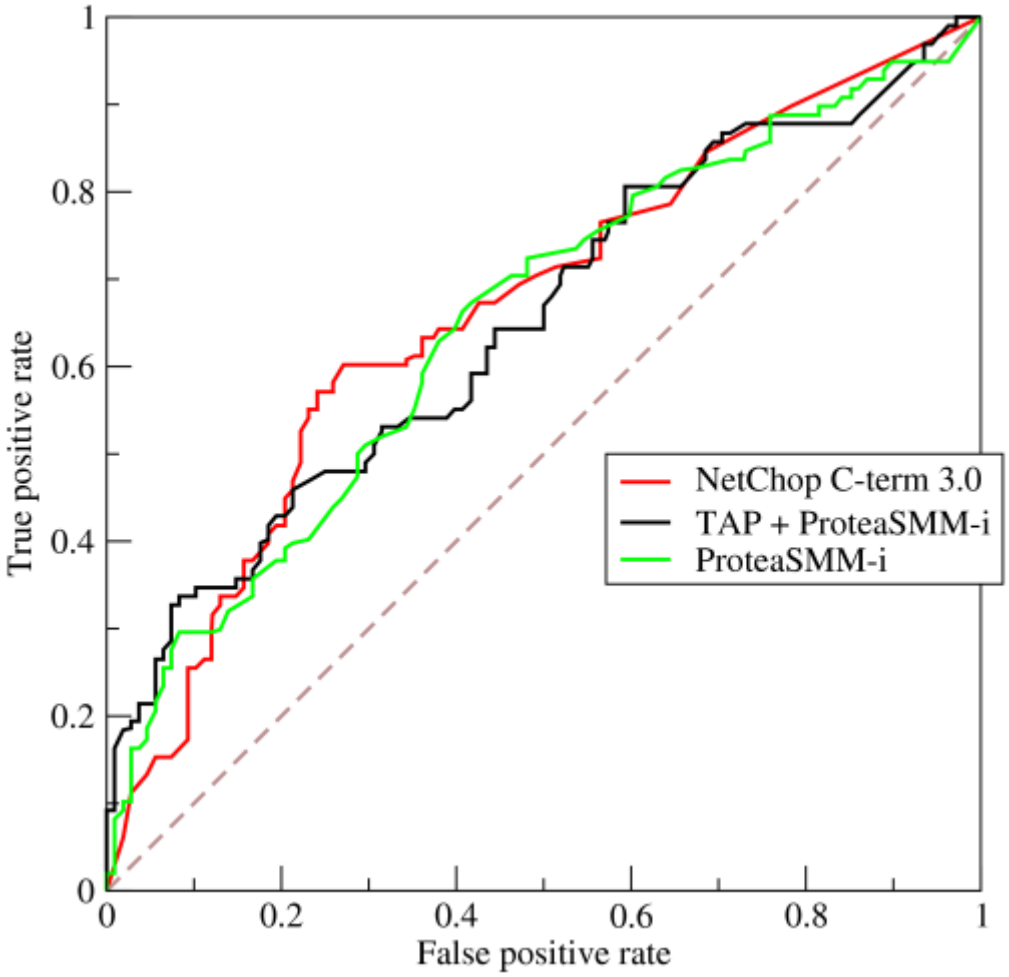

Рисунок 2.1 – Приклад ROC-кривої

Такі графіки також часто доповнюються прямою $y = x$.

Для ідеального класифікатора графік ROC-кривої проходить через верхній лівий кут, де частка істиннопозитивних та істиннонегативних випадків становить 100% або 1,0 (ідеальна чутливість), а частка хибнопозитивних та хибнонегативних прикладів дорівнює нулю. Тому чим ближче крива до верхнього лівого кута, тим вища передбачувальна здатність моделі. І навпаки, чим менше вигин кривої, і чим ближче вона розташована до діагональної прямої, тим менш ефективна модель. Діагональна лінія відповідає «даремному» класифікатору, тобто така модель абсолютно не спроможна розрізнити два класи.

При візуальній оцінці ROC-кривих розташування їх відносно одна одної вказує на їх порівняльну ефективність. Крива, розташована вище і лівіше, свідчить про більшу передбачувальну здатність моделі. Так, на рис. 2.2 дві ROC-криві суміщені на одному графіку. Видно, що модель "A" краща.

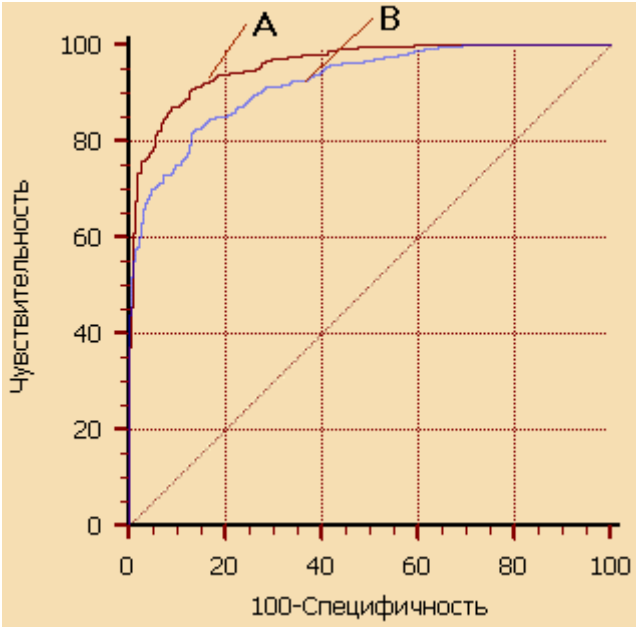

Рисунок 2.2 – Порівняння двох моделей на основі їх ROC-кривих

2.3.4. Показник AUC

Для кращого порівняння двох моделей бінарної класифікації на основі ROC-аналізу прийнято використовувати такий показник як площа під ROC-кривою, який також називають AUC (Area Under Curve).

З математичної точки зору значення AUC рівне ймовірності того, що класифікатор дасть більшу оцінку для випадково вибраного позитивного прикладу, ніж для випадково вибраного негативного випадку, якщо вважати що класифікатор і повинен давати більшу оцінку для позитивних прикладів.

Теоретично значення AUC змінюються між 0 та 1, причому значення «1» відповідає ідеальному класифікатору (недосяжному на практиці), а значення «0,5» – непотрібному класифікатору, який не є кращим за просто вгадування результату за допомогою підкидання монетки. Значення менші за «0,5» означають, що класифікатор працює навпаки, і що результати його роботи тільки покращаться, якщо вважати, що класифікатор на виході дає ймовірність прикладу бути негативним, а не позитивним. Можна також навести експертну шкалу для різних значень AUC, таку як у табл. 2.3.

Таблиця 2.3 – Експертна шкала значень AUC

Інтервал AUC	Якість моделі
0,9 – 1,0	Відмінна
0,8 – 0,9	Душе хороша
0,7 – 0,8	Хороша
0,6 – 0,7	Середня
0,5 – 0,6	Незадовільна

На практиці порахувати площу під ROC-кривою можна чисельно за допомогою методу трапецій:

$$AUC = \int f(x)dx = \frac{1}{2}\sum_{i}(X_{i+1} - X_i)(Y_{i+1} + Y_i).$$

2.3.5. Визначення порогу відсікання

Як вже було зазначено, ідеальний класифікатор має ROC-криву, що проходить через верхній лівий кут, але на практиці такий результат є

недосяжним, тому і неможливо підібрати такий поріг відсікання T, який забезпечить стовідсоткову правильність передбачень моделі.

Отже, неможливо побудувати модель, що одночасно має стовідсоткову чутливість і специфічність. Крім того, змінюючи поріг відсікання неможливо водночас повисити і чутливість, і специфічність. Тому завжди шукається деякий компроміс між цими двома величинами. Для цього вводяться певні критерії, які ми хочемо бачити виконаними в нашій моделі. Такими критеріями можуть бути:

- Обмеження на мінімальну чутливість (або специфічність) моделі. Наприклад, якщо необхідно забезпечити чутливість тесту не менше 0,8. В такому разі оптимальним буде поріг відсікання, що надає найбільше значення специфічності при одночасному дотриманні чутливості не менше 0,8. Оскільки ряд дискретний, то це можна зробити простим перебором усіх можливих значень.
- Вимога максимальної сумарної чутливості і специфічності моделі.
- Вимога балансу між чутливістю і специфічністю.

Існують також інші підходи визначення порогу відсікання, коли помилкам 1-го і 2-го другого родів призначаються різні ваги, і задача зводиться до мінімізації «загальної ваги» помилок. Але сама задача пошуку таких ваг є часто нетривіальною або навіть невирішуваною задачею.

2.4. Висновки до розділу 2

На сьогоднішній день у розпорядженні кожного аналітика є близько дюжини різних підходів для обробки даних і, зокрема, передбачення, таких

як регресія, нейронні мережі, дерева рішень, генетичні алгоритми, еволюційне програмування, нечітка логіка тощо; кожен з яких в свою чергу розділяється на десятки методів. Так, лінійна регресія в залежності від властивостей вхідних даних та від того, яку інформацію ми хочемо отримати на виході, ділиться на:

- Класичну лінійну регресію, якщо залежна змінна розподілена нормально.
- Узагальнену лінійну регресію, якщо залежна змінна розподілена за законом, що належить сімейству експоненційних.
- Логістичну регресію, якщо нам необхідно, щоб залежна змінна була бінарною або категоріальною.
- Пуассонівську регресію, якщо ми хочемо, щоб залежна змінна підраховувала кількість певних подій за встановлений проміжок часу.

Через таке різноманіття методів необхідно добре знати і розбиратися в них, для того щоб досягати оптимальних результатів ефективним шляхом.

3. АНАЛІЗ ДАНИХ ТА ПОБУДОВА МОДЕЛІ

3.1. Аналіз вибірки та попередня обробка даних

Вибірка для цього дослідження була зібрана безпосередньо з бази даних однієї з українських колекторських компаній. Вона включає у себе багато різних даних та результатів роботи з 240230 кредитними договорами за останні п’ять років. Серед них:

- ІПН (РНОКПП);
- ПІБ;
- кількість днів прострочки (DPD);
- дата виникнення прострочки;
- дата підписання кредитного договору;
- сума виданого кредиту;
- валюта кредиту;
- дата передачі договору у роботу колекторської компанії;
- усі складові суми боргу у гривнях;
- банк, у якому брали кредит;
- усі платежі (сума і дата) по цим договорам.

Слід також зазначити, що з ІПН боржника можна також витягнути такі дані як дата народження та стать позичальника. Дата народження кожного громадянина Україні закодована у перших п’яти цифрах Реєстраційного номеру облікової картки платника податків. А саме, перші п’ять цифр є кількістю днів, що пройшли з 31 грудня 1899 року до дня народження громадянина. Розрахувати це можна, наприклад, у програмі Excel за допомогою такої формули:

=”01.01.1900”+ЛЕВСИМВ(A2;5),

де A2 – це ІПН.

Стать кожного зареєстрованого платника податків можна з’ясувати за передостанньою цифрою його ІПН – якщо вона парна, то стать жіноча, якщо непарна – то чоловіча. Знову ж таки, витягнути цю інформацію у окремий стовпець можна у Excel за допомогою такої формули:

=ЕСЛИ(ЕЧЁТН(ЛЕВСИМВ(A2;9));0;1),

де A2 – це ІПН. Ця формула повертає "1", якщо клієнт є чоловіком, і "0" – якщо жінкою.

Для побудови регресії були відібрані такі змінні:

- PaysSum – сума усіх платежів позичальника за час обробки справи у колекторській компанії;
- IsMale – факторизована змінна, що приймає значення "1", якщо позичальник – чоловік, і "0" – якщо жінка;
- Age – вік користувача у роках;
- IsATO – факторизована змінна, що визначає, чи проживає позичальник на території проведення АТО або в Криму;
- BodyDelay – сума простроченого тіла кредиту на момент надання договору у роботу в колекторську компанію;
- BodyBalance - сума непростроченого тіла кредиту на момент надання договору у роботу в колекторську компанію;
- Debt – загальна сума, що має сплатити клієнт, щоб закрити кредит;
- DPD – кількість днів прострочення платежу на момент надання договору у роботу в колекторську компанію;
- DaysInWork – кількість днів опрацювання договору у колекторській компанії.

Зрозуміло, що PaysSum і DaysInWork – це дві змінні, що ми не можемо знати при оцінювання портфелю. PaysSum – це величина, яку мі маємо оцінити, а DaysInWork – параметр, який користувач має сам підставити у формулу, щоб зрозуміти, скільки заплатить позичальник за такий проміжок часу. Тому логічно за залежну змінну обрати відношення PaysSum/ DaysInWork, тобто суму платежів на один день роботи колекторської компанії.

Розглянемо тепер кожну незалежну змінну окремо. Для початку візьмемо DPD. Побудуємо точкову діаграму для ілюстрації залежності між PaysSum/ DaysInWork та DPD. Вона представлена на рис. 3.1.

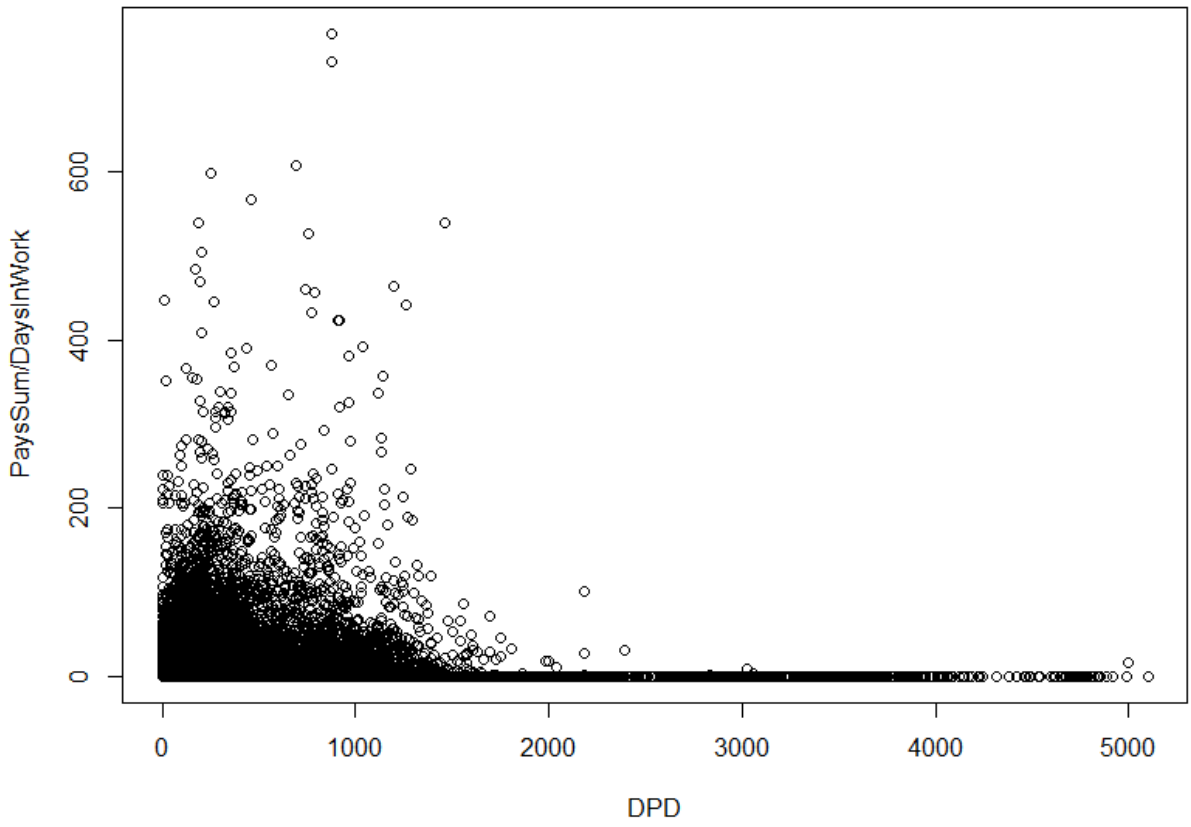

Рисунок 3.1 – Точкова діаграма залежності між PaysSum/DaysInWork і DPD

Ми бачимо, що графік фактично перестає змінюватись, якщо DPD перевищує 2000. Крім того, таких договорів доволі мало – усього 1020, тому вибірка стане лише кращою, якщо ми видалимо ці спостереження. Зрозуміло, що йдучи на такий крок ми позбавляємо цю вибірку права представляти такі договори, але якщо нам необхідно буде оцінювати портфелі з договорами з DPD більше 2000, можна без оцінювання сміливо припускати, що по цим договорам платежів не буде. Крім того, було помічено, що вибірка має 4 договори з від'ємним DPD, що, очевидно, не може бути правдою. Ці договори також треба вилучити. Така ж сама точкова діаграма але після вилучення зазначених спостережень зображена на рис. 3.2.

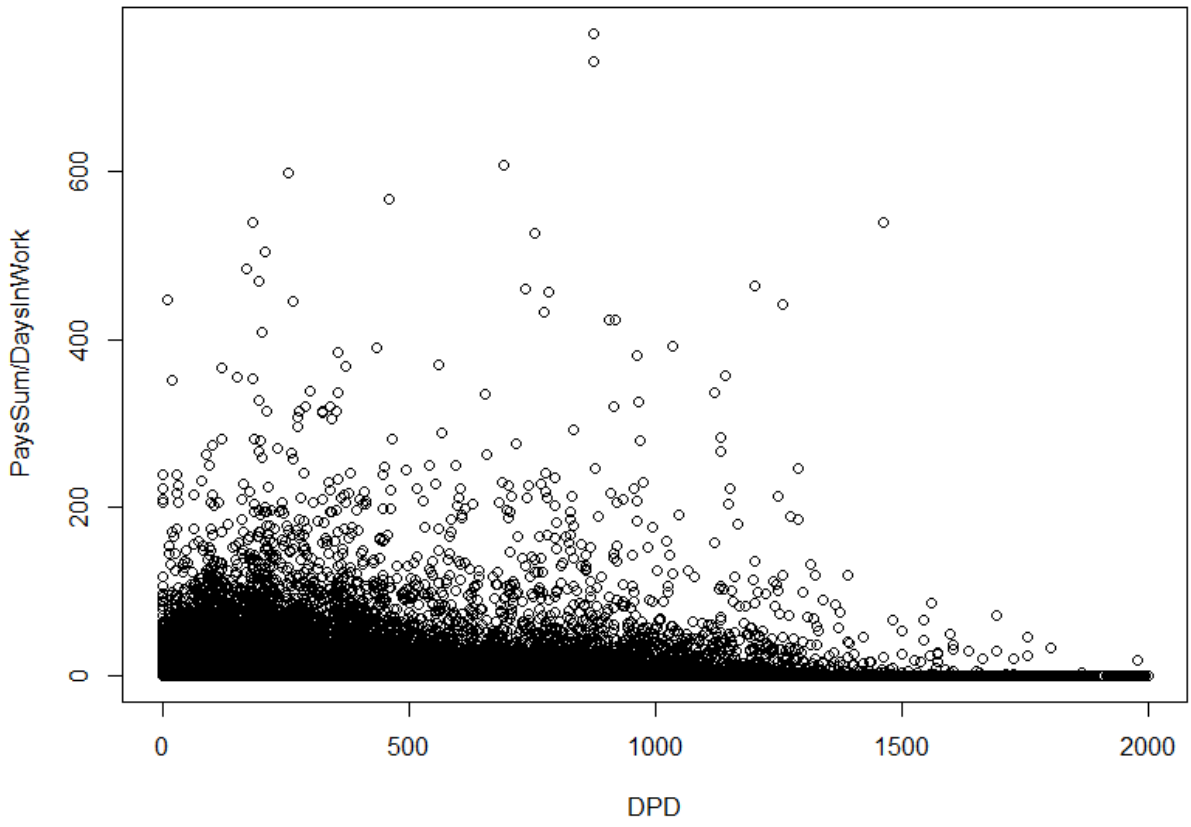

Рисунок 3.2 – Точкова діаграма залежності між PaysSum/DaysInWork і DPD після вилучення спостережень

Через величезну кількість точок на діаграмі важко простежити чітку залежність або тренд, але, загалом, можна помітити лінійну залежність, при якій зі зростанням DPD спадає сума платежів.

Далі розглянемо змінну BodyDelay. Відповідна точкова діаграма зображена на рис. 3.3.

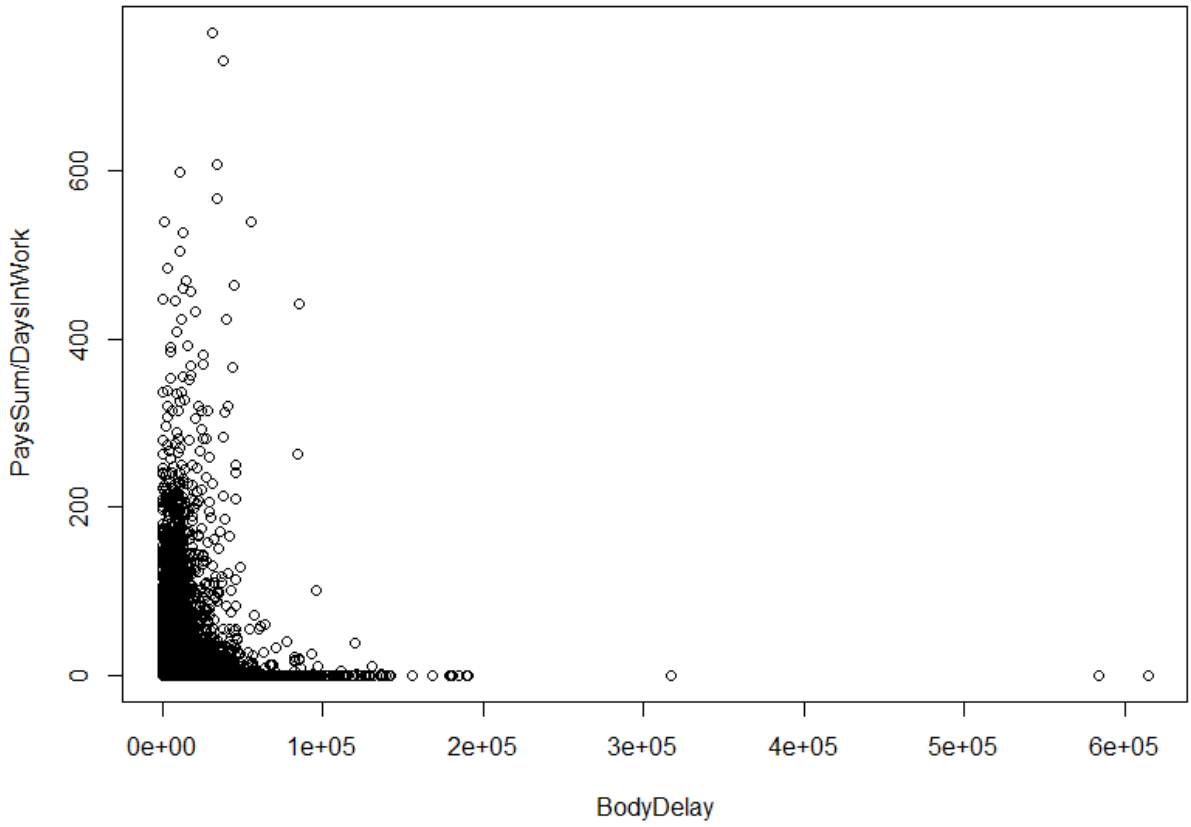

Рисунок 3.3 – Точкова діаграма залежності між PaysSum/DaysInWork і BodyDelay

Користуючись тією ж самою логікою, що й зі змінною DPD, вилучимо із дослідження усі спостереження з сумою простроченого тіла більше 50000. Таким чином ми отримуємо точкову діаграму зображену на рис. 3.4.

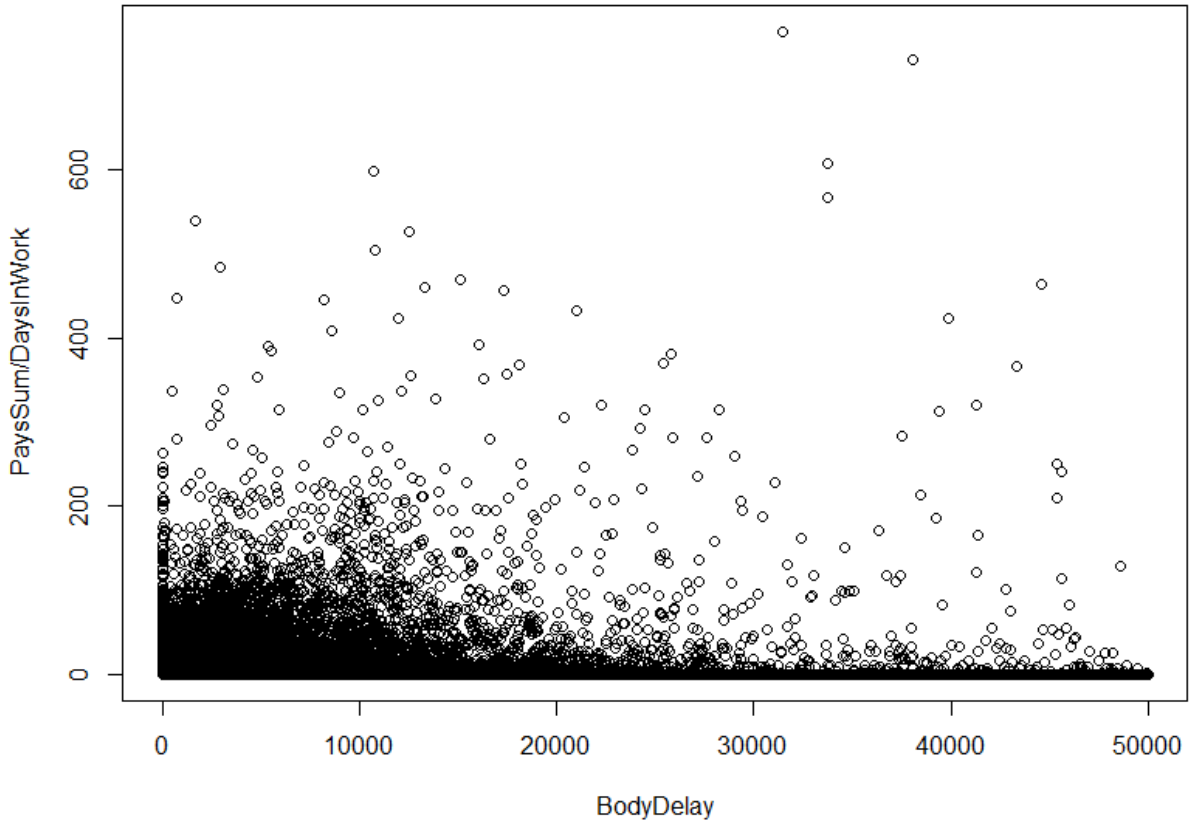

Рисунок 3.4 – Точкова діаграма залежності між PaysSum/DaysInWork і BodyDelay після вилучення спостережень

Знову ж таки, можна помітити ту ж саму обернено-пропорційну залежність, що й з DPD.

Розглянемо тепер BodyBalance. Одразу видалимо усі спостереження з сумою непрострочeного тіла більше 50000 і зобразимо отриману діаграму на рис. 3.5.

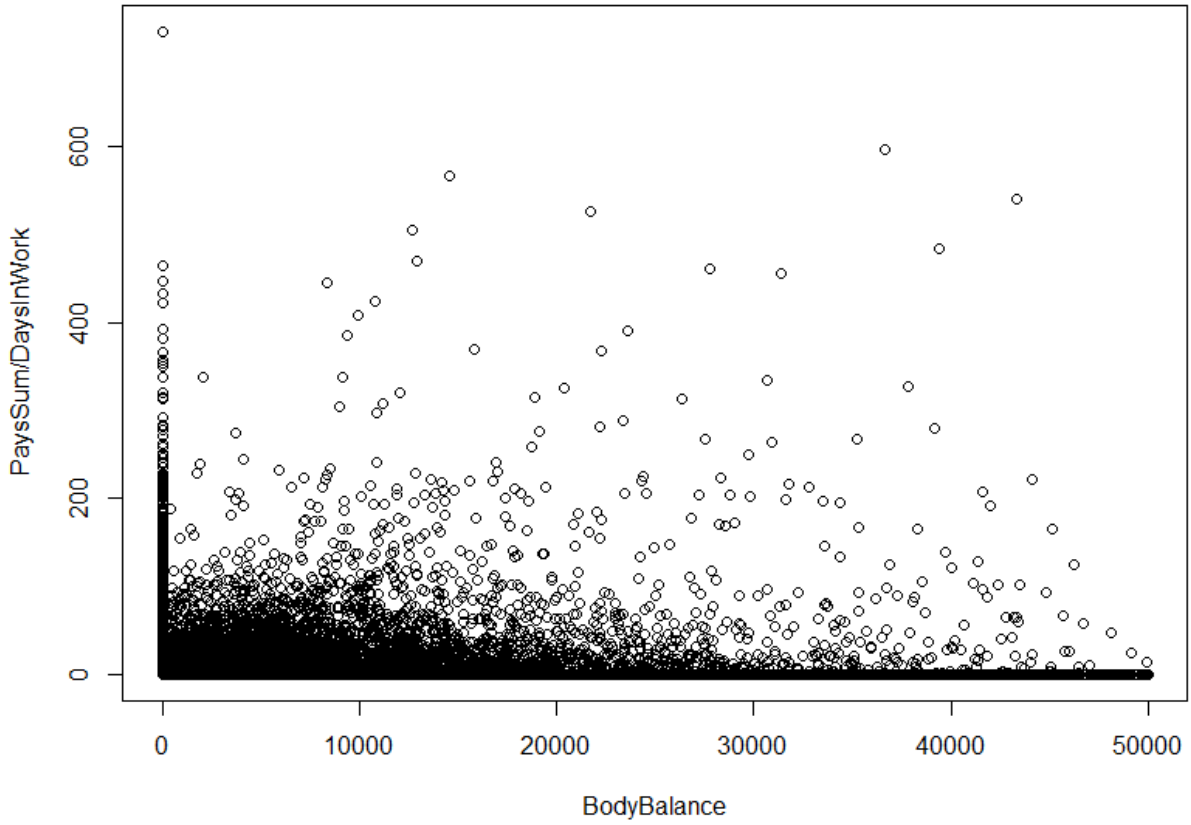

Рисунок 3.5 – Точкова діаграма залежності між PaysSum/DaysInWork і BodyBalance після вилучення спостережень

Бачимо ту саму залежність. Також слід зазначити, що через неповноту даних багато значень BodyBalance є нульовими, що може негативно вплинути на якість побудованої на таких даних регресії.

Далі на черзі змінна Debt. Після видалення усіх спостережень із загальною сумою боргу менше 100000 отримаємо діаграму зображену на рис. 3.6.

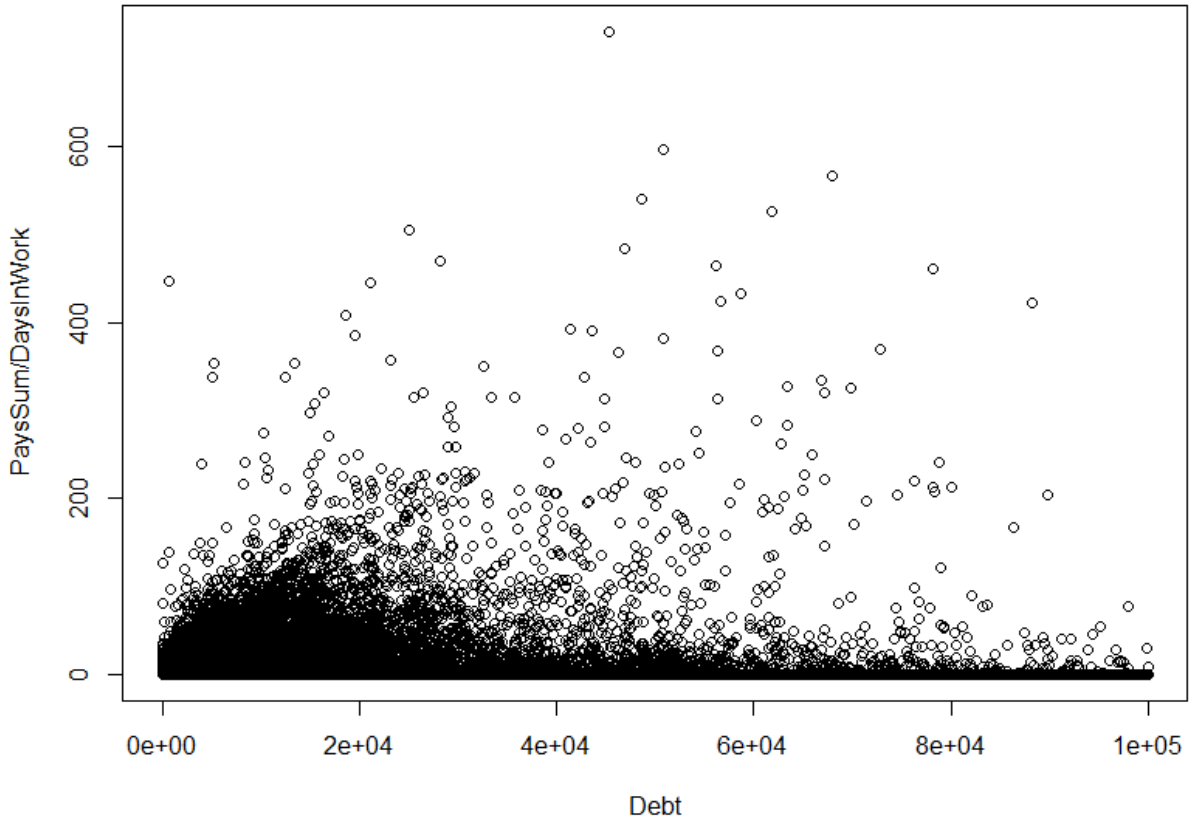

Рисунок 3.6 – Точкова діаграма залежності між PaysSum/DaysInWork і Debt після вилучення спостережень

Тут, нарешті, можна помітити дещо цікаве: залежність не лінійна. До певного моменту залежна змінна зростає при зростанні суми боргу і лише потім починає спадати. І це має сенс: позичальнику значно легше і він набагато охочіше сплатить невелику суму і погасить заборгованість, в той час як велику суму психологічно складно навіть почати сплачувати.

Розглянемо змінну Age. Відповідна точкова діаграма зображена рис. 3.7.

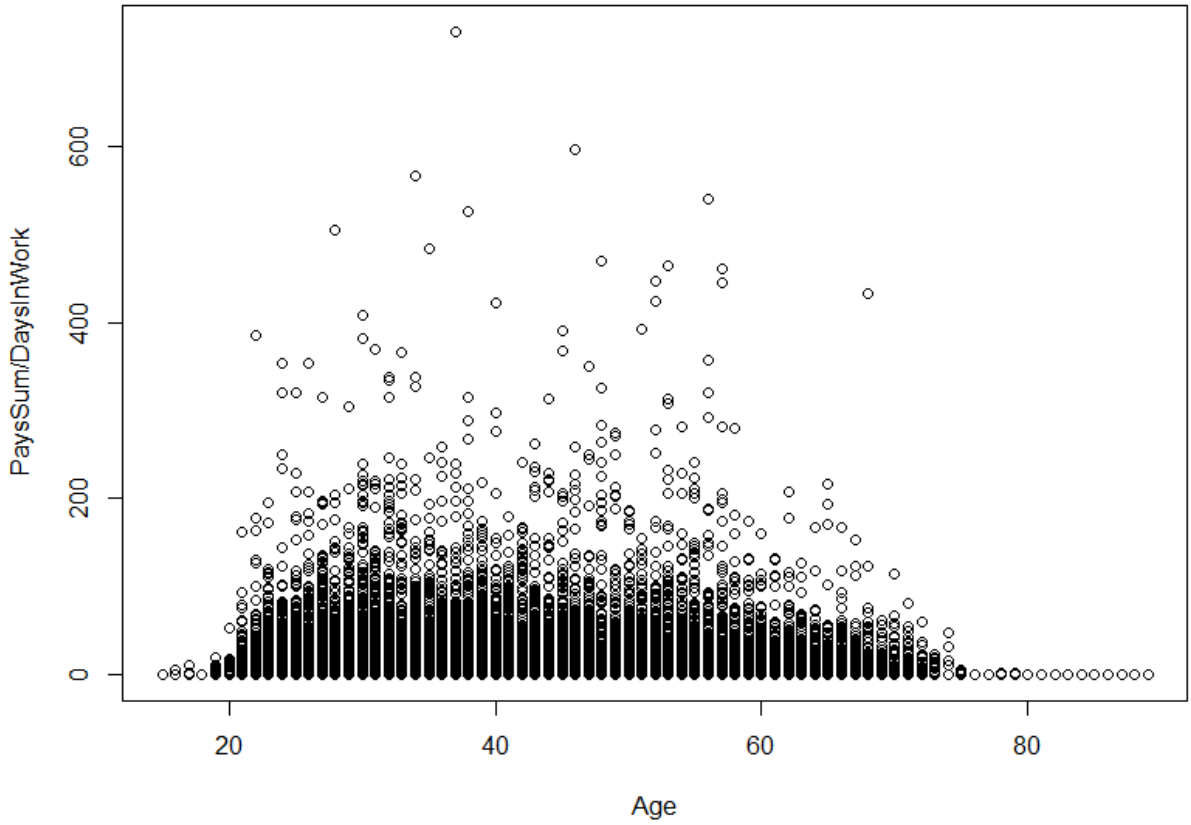

Рисунок 3.7 – Точкова діаграма залежності між PaysSum/DaysInWork і Age

Залежність видається такою ж самою, що і для Debt. Тобто до якогось віку позичальники більш схильні сплачувати зі взрослінням, а потім – навпаки.

Нарешті настала черга розглянути дві факторизовані незалежні змінні цієї вибірки: IsMale та IsATO. Зрозуміло, що будувати для них діаграми розсіювання недоречно, адже вони просто не будуть інформативними.

Усього у вибірці 105171 чоловік і 131685 жінок. При цьому чоловіки всього сплатили 31 млн гривень, а жінки – 33,9 млн. Отримаємо два відношення: 294,7 гривень в середньому сплатив один чоловік і 257,3 – одна жінка. Значення не відрізняються кардинально, тому можна зробити попередній висновок, що стать мало впливає на платоспроможність.

Змінна IsATO, мабуть, є найцікавішою особливістю цієї вибірки. Адже вона, по-перше, є характерною лише для сучасної України, і, по-друге, істотно впливає на залежну змінну, у чому ми зараз переконаємось. Усього у вибірці 178280 позичальників, що проживають на мирній Україні, і 58576 – у зоні АТО і Криму. При цьому на території мирної України було сплачено 58,7 млн гривень, а у зоні АТО і Криму – 6,1 млн. Таким чином у першому випадку на одного позичальника припадає 329,5 гривень сплаченого боргу, а у другому – лише 104,8 гривень. Різниця між показниками дуже істотна, що й не дивно, адже колекторським компаніям значно складніше проводити роботу у містах і селах, що тимчасово фактично є непідконтрольними українському законодавству.

Також слід звернути увагу на значення змінної DaysInWork цієї вибірки. Вона змінюється у межах між двома днями аж до двох років. Але, зрозуміло, що недоречно включати у вибірку спостереження з дуже малими DaysInWork, адже колекторська компанія ще не встигла попрацювати з цими позичальниками. Тому з вибірки також слід виключити усі договори, по яким робота проводилась менше місяця.

3.2. Побудова множинної регресії

Далі для усіх операцій буде використовуватися мова програмування R разом з її графічним інтерфейсом RStudio.

Для початку побудуємо множинну регресію на основі усіх незалежних змінних. Для цього потрібно розділити усю вибірку на навчальну та тестову. Виберемо співвідношення першого до другого як 75% до 25%. Далі пригадаємо, що незалежні змінні Debt і BodyDelay мали не лінійну залежність, а принаймні квадратичну. Тому включимо у регресію також

другу та третю степені кожної з цих змінних. У результаті термінал R видає результат стосовно порахованих коефіцієнтів, представлений на рис. 3.8.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.916e+00  1.215e-01  32.228  < 2e-16 ***
IsMale1         9.097e-02  5.680e-02   1.602   0.1093
Age            -1.328e-02  2.287e-03  -5.806 6.39e-09 ***
IsATO1         -1.256e+00  6.815e-02 -18.436  < 2e-16 ***
BodyDelay      -1.317e-04  2.276e-05  -5.789 7.08e-09 ***
BodyBalance     2.706e-04  7.878e-06  34.354  < 2e-16 ***
Debt            1.017e-04  1.199e-05   8.480  < 2e-16 ***
DPD            -2.913e-03  7.266e-05 -40.094  < 2e-16 ***
I(Debt^2)      -2.854e-09  3.536e-10  -8.071 7.02e-16 ***
I(BodyDelay^2)  7.233e-09  1.424e-09   5.078 3.81e-07 ***
I(Debt^3)       1.615e-14  2.909e-15   5.552 2.84e-08 ***
I(BodyDelay^3) -5.845e-14  2.414e-14  -2.421   0.0155 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 3.8 – Вивід терміналу R

По значенню стовпця `Pr(>|t|)` можна судити про значущість відповідної змінної для регресії – чим воно менше, тим більше ця змінна зв’язана зі значеннями залежної змінної. Із цього результату видно, що включення другої і третьої степенів змінних Debt і BodyDelay до регресії було доречним рішенням. У той же час, змінна IsMale фактично не впливає на значення PaysSum/DaysInWork, що й передбачалось у попередньому пункті.

Оскільки для нашої задачі нам достатньо просто зробити адекватний прогноз, а не побудувати модель, яка буде в точності відповідати дійсності, то нам не має сенсу виключати з регресії змінні, якщо тільки це не покращить прогнозувальну спроможність моделі.

Цікаво також подивитися на те, якою виявилась залежність між сумою платежів за день і загальною сумою боргу. Для цього вручну побудуємо ряд зі значеннями абсцис x від 0 до 100000 та знайдемо відповідні значення ординат y за формулою:

$$y = 1{,}016946 \cdot 10^{-4}\, x - 2{,}853815 \cdot 10^{-9}\, x^2 + 1{,}614844 \cdot 10^{-14}\, x^3.$$

Отриманий графік зображено на рис. 3.9.

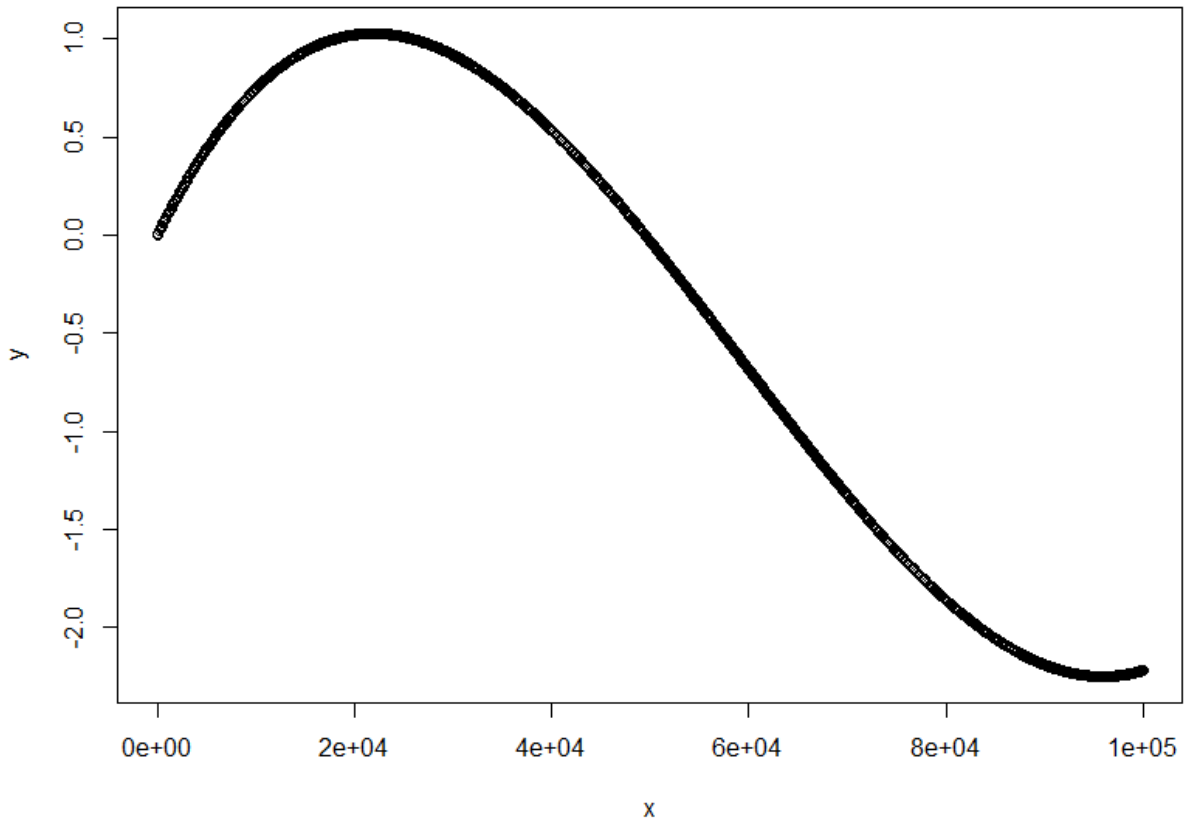

Рисунок 3.9 – Гіпотетична залежність між PaysSum/DaysInWork і Debt

Отриманий графік підтверджує попередні спостереження – до певного моменту платежі зростають при зростанні суми боргу, але потім починають спадати.

Схожа ситуація має місце і зі змінною BodyDelay. У цьому випадку залежність має таку формулу:

$$y = -1{,}317439 \cdot 10^{-4}\, x + 7.233275 \cdot 10^{-9}\, x^2 - 5.845272 \cdot 10^{-14}\, x^3.$$

Графік цієї функції зображено на рис. 3.10.

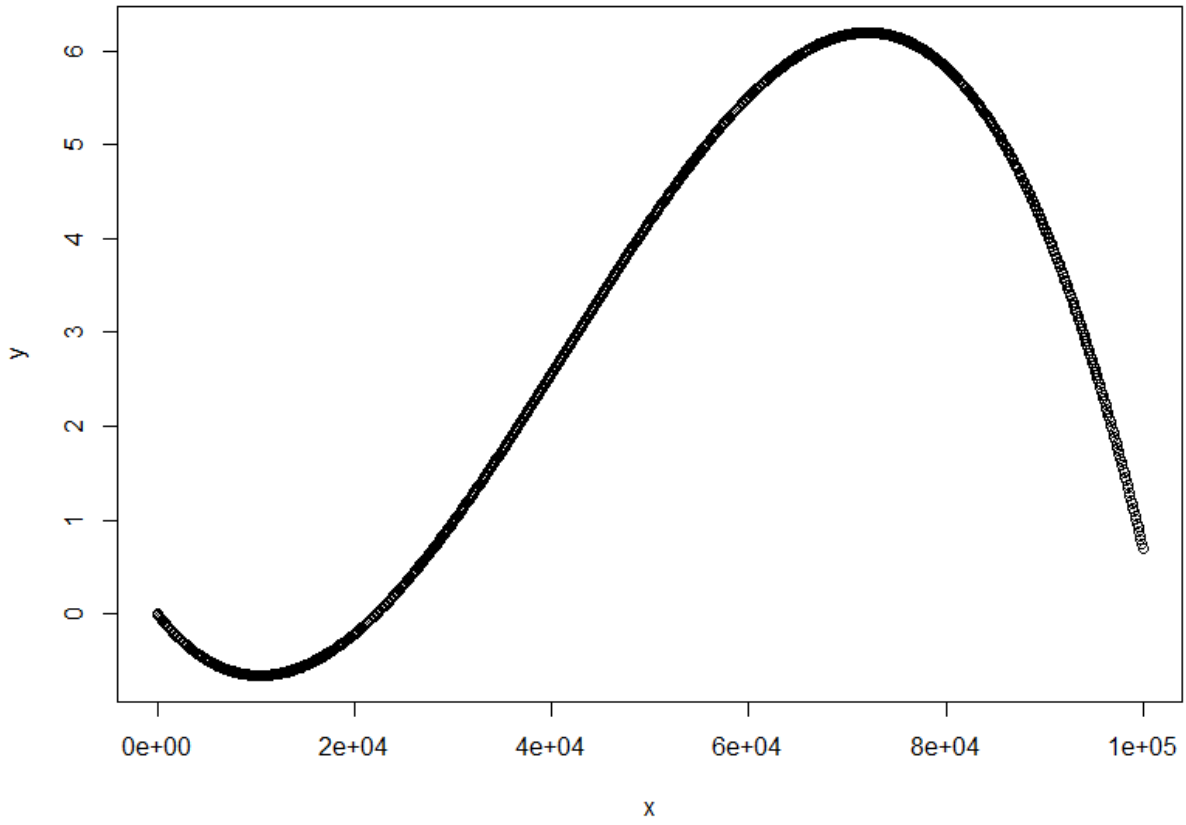

Рисунок 3.10 – Гіпотетична залежність між PaysSum/DaysInWork і BodyDelay

Тут також наявна подібна залежність між сумою платежів і сумою простроченого тіла.

3.3. Побудова альтернативної моделі

Окрім звичайної множинної регресії у цій роботі ми також спробуємо застосувати нестандартний підхід. Звичайна регресія прогнозує значення неперервно, тобто для позичальників, які скоріше за все не заплатять, вона

прогнозуватиме малі або навіть від’ємні значення, а для позичальників, що точно заплатять, – значно більші. У цій ситуації логічним чином з’являється дискретна дихотомічна змінна: чи заплатить клієнт взагалі хоч щось, чи ні.

Із другого розділу ми знаємо, що найкраще оцінювати і прогнозувати такі значення за допомогою логістичної регресії. У якості залежної змінної виступатиме новоутворена факторизована змінна HasPaid, що приймає значення "1", якщо боржник має ненульове значення PaysSum, і "0", якщо боржник не сплатив жодної гривні за весь час опрацювання у колекторській компанії.

Результати підгонки моделі представлені у рис. 3.11.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.086e-01  3.739e-02 -13.605  < 2e-16 ***
IsMale1        -6.291e-02  1.857e-02  -3.387 0.000707 ***
Age            -2.263e-03  7.491e-04  -3.021 0.002524 **
IsATO1         -7.610e-01  3.033e-02 -25.091  < 2e-16 ***
BodyDelay      -2.455e-04  7.633e-06 -32.160  < 2e-16 ***
BodyBalance     5.382e-05  2.368e-06  22.731  < 2e-16 ***
Debt           -1.136e-05  3.655e-06  -3.107 0.001890 **
DPD            -2.096e-03  3.048e-05 -68.763  < 2e-16 ***
I(Debt^2)       1.072e-10  1.184e-10   0.906 0.364960
I(BodyDelay^2)  1.299e-08  5.331e-10  24.374  < 2e-16 ***
I(Debt^3)      -2.697e-16  1.008e-15  -0.268 0.788950
I(BodyDelay^3) -1.717e-13  9.691e-15 -17.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 3.11 – Вивід терміналу R

Цікаво, що для цієї моделі друга і третя степені змінної Debt виявилися зайвими, в той час як стать позичальника почала доволі суттєво впливати на залежну змінну. При цьому від’ємне значення коефіцієнту при цій змінній означає зменшення шансу того, що боржник заплатить, якщо він є чоловіком. Слід зазначити, що такі результати не суперечать попереднім, в яких чоловіки сплачують більше, ніж жінки, у відношенні на одного боржника, а також тому, що коефіцієнт при змінній IsMale був додатній у моделі, побудованій у попередньому пункті. Такі спостереження приводять до

цікавого виводу: жінки більш схильні сплачувати борг але чоловіки в середньому сплачують більше.

Не слід забувати, що чисті значення, отримані після розрахунку залежної змінної за допомогою цієї моделі, не є ймовірністю клієнта сплатити, а є шансами на сплату. Для того, щоб отримати безпосередньо ймовірність, необхідно скористатись перетворенням:

$$Prob = \frac{1}{1 + e^{-Odds}}.$$

Після того як ми отримали ймовірність сплати хоча б одного платежу для кожного позичальника, ми можемо провести ROC-аналіз. Побудована ROC-крива для цієї моделі зображена на рис. 3.10.

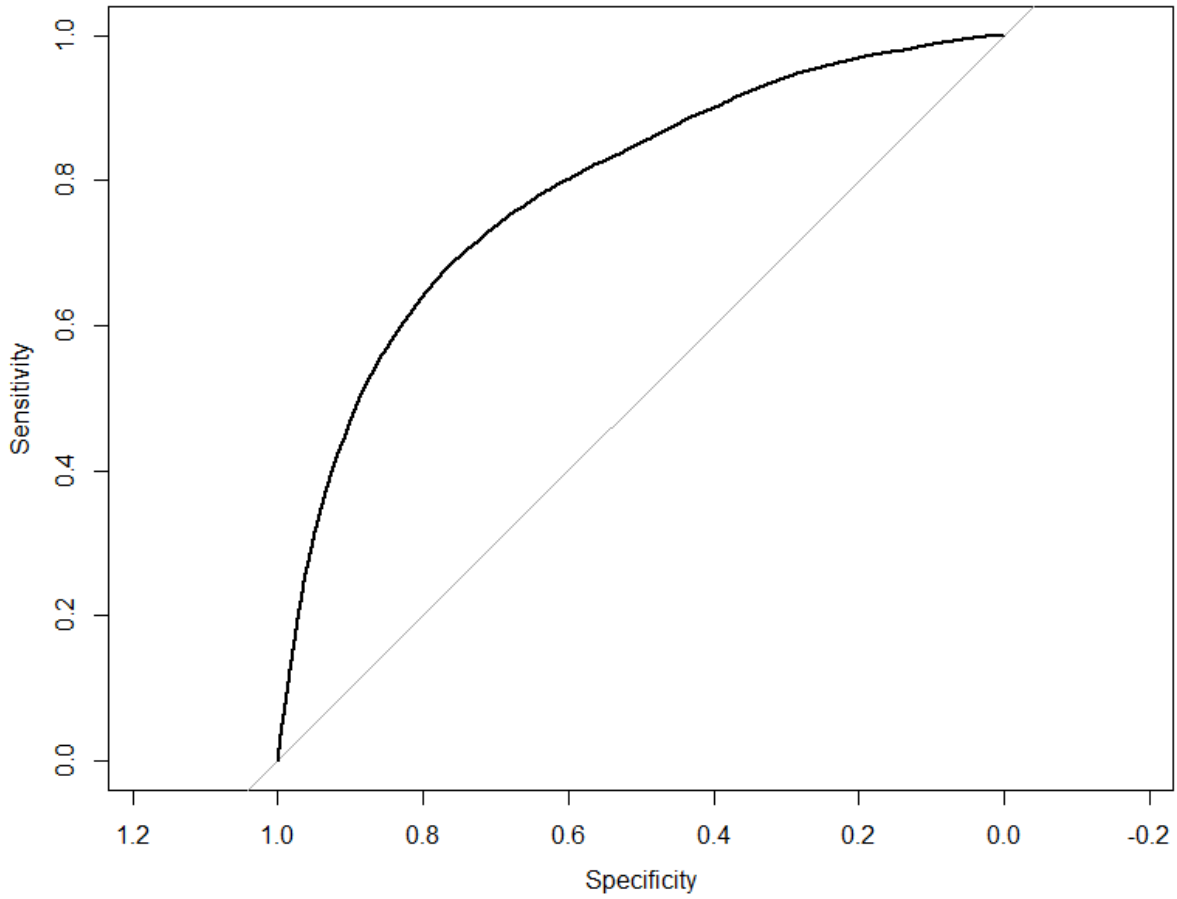

Рисунок 3.12 – ROC-крива

Значення AUC для цієї моделі становить 0,7863, що є доволі хорошим показником.

Наступним шагом є визначення порогу відсікання, тобто порогової ймовірності, при недосягненні якої ми вважатимемо, що боржник не заплатить, а при перевищенні – заплатить. Тут важливо підійти до проблеми з точки зору колекторської компанії: вона в жодному разі не хоче пропустити потенціальний дохід від клієнтів, тому можливість хибно занести клієнта до числа неплатників треба звести до мінімуму. Це означає те саме, що й зменшення можливості помилки 2-го роду, що у свою чергу означає максимізацію чутливості (Se) моделі. Тобто нам необхідно зафіксувати якесь високе значення Se (наприклад 0,9) і максимізувати значення специфічності (Sp):

$$\max_{T\in(0,1)} Sp\,, \qquad Se > 0{,}9.$$

Розв’язати її можна просто перебравши усі значення і знайти, що розв’язком цієї задачі є значення $T = 0{,}03078016$. Таблиця помилок для цієї моделі розрахована на навчальній вибірці приведена у табл. 3.1.

Табл. 3.1. Таблиця помилок.

		Передбачення	
		0	1
Істина	0	61612	93615
	1	1473	13264

Далі спробуємо використати отриману модель для побудови кращого прогнозу. Візьмемо начальну вибірку та залишимо в ній лише ті

спостереження, що пройшли поріг T. На основі цієї обрізаної вибірки підгонимо нову регресійну модель. Її коефіцієнти мають показники, наведені у рис. 3.13.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.746e+00  1.780e-01  26.672  < 2e-16 ***
IsMale1         4.991e-02  8.288e-02   0.602   0.5471
Age            -1.639e-02  3.341e-03  -4.904 9.40e-07 ***
IsATO1         -2.082e+00  1.196e-01 -17.418  < 2e-16 ***
BodyDelay      -2.574e-04  3.257e-05  -7.903 2.75e-15 ***
BodyBalance     3.360e-04  1.145e-05  29.348  < 2e-16 ***
Debt            1.291e-04  1.657e-05   7.789 6.84e-15 ***
DPD            -4.500e-03  1.389e-04 -32.397  < 2e-16 ***
I(Debt^2)      -3.871e-09  5.277e-10  -7.336 2.22e-13 ***
I(BodyDelay^2)  1.256e-08  2.255e-09   5.569 2.57e-08 ***
I(Debt^3)       2.236e-14  4.413e-15   5.066 4.08e-07 ***
I(BodyDelay^3) -9.592e-14  4.051e-14  -2.368   0.0179 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рисунок 3.13 – Вивід терміналу R

Бачимо, що результати, в цілому, схожі з моделлю отриманою у попередньому пункті.

Використовуватися ця модель буде наступним чином: спочатку логістична регресія визначає, які клієнти із тестової вибірки взагалі не будуть платити, а на решті застосовується побудована щойно множинна регресія.

3.4. Порівняння моделей

Для порівняння моделей потрібно використати кожну побудовану модель для прогнозування на тестовій вибірці. При цьому потрібно звернути увагу на те, що моделі можуть давати від’ємні значення. Зрозуміло, що їх потрібно заміняти нулями. Зробивши це, було виявлено дуже цікавий факт стосовно другої моделі. Якщо спочатку використати множинну регресію на тестовій вибірці, то можна помітити, що отримані прогнозовані значення є

від’ємними як раз на тих значеннях, які б ми відсікли логістичною регресією. Природа цього факту не є ясною, і можливо це лише співпадіння. В теорії така властивість може означати, що нам взагалі не потрібно нічого відсікати логістичною регресією, тому що побудована на обрізаній вибірці множинна регресія зробить це сама, адже ми заміняємо на нуль усі спрогнозовані від’ємні значення.

Наступним кроком необхідно обрати критерій порівняння. Оскільки, як вже було сказано, для моделей прогнозування важливою є лише точність прогнозування і ніякі інші показники, то за критерій порівняння можна обрати просто суму квадратів залишків моделей.

Для першої моделі вона становить: 167805520499, а для другої – 166977918033. Різниця, хоча й невелика, але на користь другої, альтернативної, моделі.

Для інтересу також можна привести порівняння реальної загальної суми зборів з двома прогнозованими (табл. 3.2).

Таблиця 3.2 – Порівняння прогнозованих загальних сум зборів

Модель	Загальна сума зборів, грн
Реальна сума	16598728
Стандартна модель	18415426
Альтернативна	15763249

Чітко видно, що перша модель значно завищує збори боржників, у той час як друга – лише трохи їх занижує.

3.5. Висновки до розділу 3

Задачею цього розділу була побудова регресійної моделі прогнозування платежів боржників. У ході цієї розробки були виявлені деякі цікаві та, навіть, неочікувані закономірності щодо залежності між сумою платежів та статтю, сумою боргу. Крім того, особливість предметної області і значення, що прогнозується, дали можливість розробити кардинально інший підхід до вирішення задачі, що в результаті вийшов кращим за стандартний.

4. СТАРТАП-ПРОЕКТ «СМАРТ ДЕБТ»

Суть цієї магістерської дисертації полягає у розробці і програмній реалізації методу оцінювання вартості кредитного портфелю проблемної заборгованості. Такий програмний продукт можна як продати вже існуючим банкам чи колекторським компаніям, так і, залучившись підтримкою інвесторів, заснувати власну колекторську компанію. У рамках цього розділу будемо вважати, що ми відкриваємо власну колекторську компанію під назвою «Смарт Дебт».

4.1. Опис ідеї проекту

У табл. 4.1 надано зміст ідеї (що пропонується), можливі напрямки застосування та основні вигоди, що може отримати користувач товару.

Таблиця 4.1 – Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Створення власної колекторської компанії «Смарт Дебт» на основі розробленого програмного продукту	1. недороге та точне оцінювання портфелів	Економія грошей та правильні рішення щодо покупки портфелів
	2. прогнозування зборів по портфелю	Ефективне керування фінансовими активами компанії
	3. високоефективна робота з боржниками	Більший відсоток повернення вкладів

Виділимо такі техніко-економічні характеристики ідеї:

1) можливість та точність оцінювання портфелів;
2) наявність кваліфікованих менеджерів;
3) наявність власного кол-центру.

Для порівняння цього проекту з іншими представниками на ринку, у якості конкурентів виберемо такі три колекторські компанії:

1) Вердикт Консалтинг;
2) Долгофф;
3) Укрборг.

Таблиця 4.2 – Визначення сильних, слабких та нейтральних характеристик ідеї проекту

№	Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів				W (слабка сторона)	N (нейтральна сторона)	S (сильна сторона)
		Смарт Дебт	Вердикт	Долгофф	Укрборг			
1.	оцінювання портфелів	+	-	+	+			+
2.	менеджери	-	+	-	+		+	
3.	кол-центр	-	+	-	+	+		

4.2. Технологічний аудит ідеї проекту

Визначення технологічної здійсненності ідеї проекту передбачає аналіз складових, наведених у табл. 4.3.

Таблиця 4.3 – Технологічна здійсненність ідеї проекту

№ п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій
1	Облаштування офісу	Оренда чи покупка (у кредит)	Наявні	+
2	Покупка портфелів заборгованості	Оцінювання портфелів проводиться за допомогою розробленого програмного продукту; гроші беруться або у інвесторів, або в довгостроковий кредит	Наявні	+
3	Обробка портфелів кол-центром	Покупка послуг кол-центру сторонньої компанії або створення власного з нуля	Наявні	+
Обрана технологія реалізації ідеї проекту: офіс будемо орендувати; для покупки портфелів залучимо інвесторів; користуватимемось послугами стороннього кол-центру.				

4.3. Аналіз ринкових можливостей запуску стартап-проекту

Спочатку проводиться аналіз попиту: наявність попиту, обсяг, динаміка розвитку ринку (табл. 4.4).

Таблиця 4.4. Попередня характеристика потенційного ринку стартап-проекту

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	10
2	Загальний обсяг продаж, грн	2 млрд
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу	
5	Специфічні вимоги до стандартизації та сертифікації	Потрібна сертифікація
6	Середня норма рентабельності в галузі (або по ринку), %	25

Надалі визначаються потенційні групи клієнтів, їх характеристики, та формується орієнтовний перелік вимог до товару для кожної групи (табл. 4.5).

Таблиця 4.5 – Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Потреба у стягування боргів з клієнтів	Банки та інші кредитні установи України		- ефективність; - відповідальність; - толерантність до боржників.

Після визначення потенційних груп клієнтів проводиться аналіз ринкового середовища: складаються таблиці факторів, що сприяють ринковому впровадженню проекту, та факторів, що йому перешкоджають (табл. 4.6 і 4.7). Фактори в таблиці подані в порядку зменшення значущості.

Таблиця 4.6 – Фактори загроз

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Введення в оману банком	Банк чи інша установа може надати недостовірні дані в описі портфелю	Оскарження у суді
2	Хибне оцінювання портфелю	При оцінці портфелю можуть бути не враховані важливі фактори, які суттєво скажуться на похибці	Нічого не поробиш, наступного разу такого не повториться
3	Людська помилка при роботі з портфелем	Ненавмисна неправильна обробка інформації в процесі роботи з портфелем	Штрафування або звільнення працівника, що допустився помилки

Таблиця 4.7 – Фактори можливостей

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Ініціативність та креативність працівників	Розробка менеджером нової стратегії роботи з клієнтами	Заохочення таких працівників бонусами та преміями

Далі проводиться аналіз пропозиції: визначаються загальні риси конкуренції на ринку (табл. 4.8).

Таблиця 4.8 – Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Вказати тип конкуренції - монополія / олігополія / монополістична / чиста	Чиста	Гарні перспективи розвитку
2. За рівнем конкурентної боротьби - локальний / національний / …	Національний	Ведучи конкуренцію на національному рівні, компанії необхідно прикласти належні зусилля для охоплення всього національного ринку.
3. За галузевою ознакою - міжгалузева / внутрішньогалузева	Внутрішньогалузева	Необхідно зосередити зусилля на пошуку конкурентних переваг, які дозволять компанії займати стійкі конкурентні позиції на даному ринку.
4. Конкуренція за видами товарів: - товарно-родова - товарно-видова - між бажаннями	Товарно-видова	
5. За характером конкурентних переваг - цінова / нецінова	Нецінова. При виборі колекторської компанії банк звертає увагу на надійність та досвід компанії. Цінова. Для значної частки банків ціна є визначальною при виборі.	Зосередити зусилля на накопичені досвіду та отриманні солідного іміджу серед колекторських компаній
6. За інтенсивністю - марочна / не марочна	Марочна	

Після аналізу конкуренції проводиться більш детальний аналіз умов конкуренції в галузі (табл. 4.9).

Таблиця 4.9 – Аналіз конкуренції в галузі за М. Портером

	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
Складові аналізу	Навести перелік прямих конкурентів	Визначити бар'єри входження в ринок	Визначити фактори сили постачальників	Визначити фактори сили споживачів	Фактори загроз з боку замінників
Висновки:	На ринку спостерігається збільшення кількості гравців, що сприяє кращій конкуренції.	Бар'єри входу на ринок є порівняно незначними. Вартість організації бізнесу складає 1 млн. грн. Обов'язковою є сертифікація продукту.	Немає залежності від постачальників.	Клієнти значною мірою впливають на загальний попит.	Субститутів немає.

За результатами аналізу таблиці робиться висновок щодо принципової можливості роботи на ринку з огляду на конкурентну ситуацію (табл. 4.10).

Таблиця 4.10 – Обґрунтування факторів конкурентоспроможності

№ п/п	Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)
1	Науковість підходу	Метод оцінювання портфелів є обґрунтованим з математичної точки зору
2	Гарантії ефективної роботи	Висока точність прогнозів дозволяє краще управляти фінансами
3	Толерантність роботи з боржниками	Робота операторів кол-центру ретельно контролюється

За визначеними факторами конкурентоспроможності (табл. 4.10) проводиться аналіз сильних та слабких сторін стартап-проекту (табл. 4.11).

Таблиця 4.11 – Порівняльний аналіз сильних та слабких сторін «Смарт Дебт»

№ п/п	Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з «Смарт Дебт»						
			–3	–2	–1	0	+1	+2	+3
1	Науковість підходу	20	+						
2	Гарантії ефективної роботи	17			+				
3	Толерантність роботи з боржниками	13					+		

Фінальним етапом ринкового аналізу можливостей впровадження проекту є складання SWOT-аналізу (матриці аналізу сильних (Strength) та слабких (Weak) сторін, загроз (Troubles) та можливостей (Opportunities) (табл. 4.12).

Таблиця 4.12 – SWOT- аналіз стартап-проекту

Сильні сторони: науковість підходу, гарантії ефективної роботи	Слабкі сторони: відсутність досвіду, відсутність власного кол-цетру
Можливості:	Загрози: введення в оману банком, хибне оцінювання портфелю, людська помилка при роботі з портфелем

На основі SWOT-аналізу розробляються альтернативи ринкової поведінки (перелік заходів) для виведення стартап-проекту на ринок та орієнтовний оптимальний час їх ринкової реалізації з огляду на потенційні проекти конкурентів, що можуть бути виведені на ринок (див. табл. 9, аналіз потенційних конкурентів).

Визначені альтернативи аналізуються з точки зору строків та ймовірності отримання ресурсів (табл. 4.13).

Таблиця 4.13 – Альтернативи ринкового впровадження стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	Диверсифікація	Заснування різних «підкомпаній», що будуть спеціалізуватись на роботі з конкретними банками чи кредитними установами	2 місяці
2	Розвиток ринку	Вихід продукту на міжнародний ринок	6 місяців

4.4. Розроблення ринкової стратегії проекту

Розроблення ринкової стратегії першим кроком передбачає визначення стратегії охоплення ринку: опис цільових груп потенційних споживачів (табл. 4.14).

Таблиця 4.14 – Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Важкість входу у сегмент
1	Банки	середня	середній	висока	середня
2	Онлайн-кредитування	висока	середній	висока	середня
Які цільові групи обрано: компанії онлайн-кредитування.					

Для роботи в обраних сегментах ринку необхідно сформувати базову стратегію розвитку (табл. 4.15).

Таблиця 4.15 – Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
1	Спеціалізація	Зосередження на одному сегменті ринку	Науковість підходу, гарантії ефективної роботи	Стратегія спеціалізації

Наступним кроком є вибір стратегії конкурентної поведінки (табл. 4.16).

Таблиця 4.16 – Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки
1	Ні	Забиратиме існуючих у конкурентів	Усі колекторські компанії і так досить схожі одна на одну	Стратегія заняття конкурентної ніші

На основі вимог споживачів з обраних сегментів до постачальника (стартап-компанії) та до продукту (див. табл. 4.5), а також в залежності від обраної базової стратегії розвитку (табл. 4.15) та стратегії конкурентної поведінки (табл. 4.16) розробляється стратегія позиціонування (табл. 4.17). що полягає у формуванні ринкової позиції (комплексу асоціацій), за яким споживачі мають ідентифікувати проект.

Таблиця 4.17 – Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
1	ефективність; відповідальність; толерантність до боржників.	Стратегія спеціалізації	науковість підходу, гарантії ефективної роботи, толерантність роботи з боржниками	Науковість, точність, толерантність

4.5. Розроблення маркетингової програми стартап-проекту

Першим кроком є формування маркетингової концепції товару, який отримає споживач. Для цього у табл. 4.18 потрібно підсумувати результати попереднього аналізу конкурентоспроможності товару.

Таблиця 4.18. Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1	Ефективне стягування заборгованостей	Точність оцінки та прогнозу забезпечує високу ефективність	У більшості конкурентів цим речам не приділяють так багато уваги
2	Толерантність по відношенню до боржників	Суворий контроль над операторами кол-центру щодо цього	Більшість конкурентів також за цим слідкують

Надалі розробляється трирівнева маркетингова модель товару: уточнюється ідея продукту та/або послуги, його фізичні складові, особливості процесу його надання (табл. 4.19).

Таблиця 4.19 – Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові		
I. Товар за задумом	Розумний програмний продукт із високим рівнем точності оцінювання та зручним інтерфейсом.		
II. Товар у реальному виконанні	Властивості/характеристики	М/Нм	Вр/Тх /Тл/Е/Ор
	1. Функціональність 2. Швидкодія 3. Зручність 4. Зовнішній вигляд інтерфейсу	М М М Нм	Тх Тх Е Е/Ор
	Якість: продукт має відповідати міжнародним стандартам		
III. Товар із підкріпленням	Використовується реклама		
	Використовуються тимчасові знижки		
За рахунок чого потенційний товар буде захищено від копіювання: за рахунок електронних ключів та інтелектуальної власності.			

Наступним кроком є визначення цінових меж, якими необхідно керуватись при встановленні ціни на потенційний товар (остаточне визначення ціни відбувається під час фінансово-економічного аналізу проекту), яке передбачає аналіз ціни на товари-аналоги або товари субститути, а також аналіз рівня доходів цільової групи споживачів (табл. 4.20).

Таблиця 4.20 – Визначення меж встановлення ціни

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
1	Немає	Трохи вищий	Високий	від 5% до 30% від зборів

Наступним кроком є визначення оптимальної системи збуту, в межах якого приймається рішення (табл. 4.21).

Таблиця 4.21 – Формування системи збуту

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Банки надають у роботу чи повністю продають портфелі проблемної заборгованості	Робити звіти про виконану роботу по портфелях		

Останньою складової маркетингової програми є розроблення концепції маркетингових комунікацій, що спирається на попередньо обрану основу для позиціонування, визначену специфіку поведінки клієнтів (табл. 4.22).

Таблиця 4.22 – Концепція маркетингових комунікацій

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
1		Телефон, e-mail		Привернути увагу та зацікавити цільових клієнтів.	Надання знижок на перші місяці користування продуктом.

4.6. Висновки до розділу 4

У результаті опрацювання цього розділу був розроблений комерційно спроможний проект колекторської компанії. Для цього були оглянуті та запропоновані ринкові стратегії; були описані потенційні конкуренти та переваги і недоліки у порівнянні з ними; були розглянуті бар’єри входження на ринок. Даний проект доцільно розвивати та реалізовувати.

ВИСНОВКИ

В останні роки Українська економіка перебуває у стані кризи, що спричинило значне підвищення кількості проблемних кредитів. Фізичні та юридичні особи беруть гроші у банків у борг але не можуть їх повернути. Оскільки ця проблема є лише симптомом економічної нестабільності, її вирішення не розв’яже усі проблеми у країні, проте буде важливим кроком на шляху до цього.

Існує два способи зменшення частки проблемних заборгованостей: не видавати кредити завідомо неплатоспроможним клієнтам і примусово стягувати кошти з боржників. Обидва способи потребують розробки певних моделей передбачення для їхньої реалізації. Так, коли клієнт приходить у банк із наміром взяти кредит, він має спочатку заповнити ретельно підготовану анкету, що збирає інформацію про клієнта для двох конкретних цілей: однозначної ідентифікації клієнта (ПІБ, ІПН, мобільний номер тощо) та визначення його платоспроможності (наявність роботи, рівень доходів, кредитна історія, територіальна приналежність до зони АТО тощо). Потім відповіді на ці запитання разом із інформацією про кредит (загальна вартість кредиту, сума щомісячного платежу, наявність першого платежу) подаються на вхід розробленої аналітиками банку математичної моделі, що на виході видає просту відповідь – видавати кредит чи відмовити.

У свою чергу, примусове стягнення коштів з боржників є значно більш проблематичним завданням, що має багато різних підходів до його вирішення. По-перше, треба чітко визначати, які кредити взагалі вважати проблемними. Ті, що прострочені на тиждень, місяць чи три місяці? Або, можливо, брати за критерій не кількість днів прострочення, а суму простроченої заборгованості, і вважати боржниками тих, у кого вона більша за одну чи дві тисячі гривень? Кожен банк вирішує це питання окремо для

себе і діє відповідно до їхньої політики стосовно цього питання. Після визначення проблемних клієнтів, необхідно зрозуміти, чому вони не платять, та вирішити, що з ними робити. На ранніх строках прострочення усі банки самостійно намагаються вплинути на боржників за допомогою текстових повідомлень та/або дзвінків. З більш зухвалими неплатниками банк має три різні шляхи роботи:

- самостійний збір простроченої заборгованості;
- аутсорсинг збору простроченої заборгованості;
- продаж простроченої заборгованості з балансу.

Кожен з трьох варіантів потребує розробки банком математичних моделей. Так, у першому випадку, необхідно розробити CRM-систему, що буде зберігати для обробляти усю інформацію про усіх боржників та, якщо вони вийшли на прострочку, самостійно у відповідності до політики банку стосовно проблемних кредитів вирішувати, які дії необхідно робити. Для аутсорсингу необхідно визначити економічно обґрунтовану величину комісії, що буде повертати наймана компанія при надходженні коштів від боржників. І ця величина може залежати від багатьох факторів: сума платежу, наявність залогу, кількість днів прострочення, територіальна приналежність до зони АТО тощо).

Ця робота стосується третього варіанту – продажу кредитного портфелю. Очевидно, де йдеться продаж, необхідно домовитися про ціну, що задовільнила б обидві сторони. Банк має намір повністю передати усі права на проблемні кредити у руки третіх осіб – колекторської компанії. Остання, у свою чергу хоче знати, наскільки вигідною буде така інвестиція. Для цього необхідно передбачити грошовий потік (cash flow), що буде генерувати придбаний портфель (кошти стягнені з боржників мінус затрати на дії по стягненню), та на його основі розрахувати внутрішню норму прибутку (IRR). Грошовий потік, як тільки що було зазначено, складається з коштів, що

платять боржники з урахуванням затрат на роботу колекторської компанії та, звісно, суми інвестиції. У цій роботі розроблено математичну модель, що вираховує лише кошти, зібрані від боржників, тому що розрахунок витрат і підрахунок IRR не стосуються безпосередньо проблеми передбачення якості портфелю. Крім того, модель розрахована тільки на беззалогові кредити фізичних осіб, оскільки наявність залогу потребує значного ускладнення методів, а великі кредити юридичних осіб потребують індивідуального підходу, і будувати для них модель не є доцільним чи, навіть, можливим завданням.

У першому розділі на основі проведеної дослідницької роботи було описано багато методів розрахунку можливих платежів від боржників, що основані на різних математичних методах, які наразі використовуються в Україні й в світі у цій сфері. На основі цього досвіду було обрано регресійні методи як математичну основу для моделі, розробленої у цій роботі.

У другому розділі описуються теоретичні засади, використанні при розробці моделі.

Третій розділ описує багато різних процесів. Спочатку розглядається вибірка, що була доступна для цієї роботи. З вибірки виокремлюються певні поля, обрані на основі дослідження, проведеного у першому розділі, які будуть використані як незалежні змінні для множинної регресії. На основі інтуїтивного судження, що існує принципіальна різниця між боржниками, що платять хоч щось, та тими, хто ніколи нічого не заплатить, було вирішено використати логістичну регресію для розділення боржників на ці дві групи та розраховувати можливі платежі лише на основі першої групи. Після цієї підготовки обидві запропоновані моделі реалізуються у середовищі RStudio на мові програмування R. Було отримано та описано деякі цікаві та непередбачувані результати роботи цих моделей. Та загалом, прийнято висновок, що друга, альтернативна модель є кращою.

Цю роботу можна в подальшому доповнити та покращити, включивши в розрахунок витрати на роботу з портфелем колекторської компанії, що дозволить повноцінно передбачувати грошовий потік та оцінювати внутрішню норму прибутку. Крім того можливо розширити цю модель для оцінювання не тільки беззалогових але й залогових кредитів. Для цього необхідно буде дослідити, які з показників, що характеризують залог, слід включити у регресію. Також за допомогою мінімальних змін можна адаптувати розроблену модель для розрахунку вигідної для обох сторін комісії при аутсорсингу банком своїх проблемних активів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Мертвий вантаж. Чому в українських банках проблемних кредитів більше, ніж "живих". *Незалежна асоціація банків України* : веб-сайт. URL: https://nabu.ua/ua/mertviy-vantazh-chomu-v-ukrayinskih.html (дата звернення: 15.10.2018).

2. Оценка портфелей просроченной задолженности физических лиц при подготовке договоров цессии. *Банкир.Ру* : веб-сайт. URL: https://bankir.ru/publikacii/20081201/ocenka-portfelei-prosrochennoi-zadoljennosti-fizicheskih-lic-pri-podgotovke-dogovorov-cessii-1427249/ (дата звернення: 15.10.2018).

3. Хейнсворт Р., Николаенко Е., Макаренко Л. Обзор и оценка проблемных кредитов: потенциал рынка. 2010. 71 с. URL: http://www1.ifc.org/wps/wcm/connect/f33e340048fc6638b509bd849537832d/RussiaCR-NPL-SurveyReport-RU.pdf?MOD=AJPERES (дата звернення: 15.10.2018).

4. Болгар Т. М. Проблемні кредити банків як результат реалізації кредитного ризику. *Економічний нобелівський вісник*. 2014. №1. С. 50–58. URL: http://nbuv.gov.ua/UJRN/bmef_2014_1_10 (дата звернення: 15.10.2018).

5. Денисенко М. П., Домрачев М. П., Кабанов В. Г. Кредитування та ризики : навчальний посібник. Київ : ВД «Професіонал», 2008. 480 с.

6. Вовк В. Я., Хмеленко О. В. Кредитування і контроль : навчальний посібник. Київ : Знання, 2008. 463 с.

7. Купчинова О. Проблемная кредитная задолженность: подходы к определению. *Банковский вестник*. 2010. № 16. С. 42–48.

8. Осокина Т. М. Бухгалтерский учет в банках : учебный курс. Москва : МИЭМП, 2010. 148 с.

9. Кльоба В. Л. Ситуаційний центр банку як ефективний напрям удосконалення управління врегулюванням проблемної заборгованості.

Науковий вісник НЛТУ України. 2009. № 19.8. С. 240–246.

10. Шустова Е. П. «Проблемный кредит»: терминологическое содержание, критерии определения и факторы возникновения. *Вестник Алтайской академии экономики и права*. 2010. № 18. С. 155–158.

11. Кузнецов С. В. Ссудная задолженность кредитных организаций: проблемы и инструменты её урегулирования : автореф. дис. Москва : Академия народного хозяйства при Правительстве Российской Федерации, 2008. 20 с.

12. Нурзат О. А. Перспективные подходы к повышению эффективности управления проблемными кредитами в коммерческих банках : автореф. дис. Москва : Академия народного хозяйства при Правительстве Российской Федерации, 2011. 18 с.

13. Хейнсворт Р., Николаенко Е., Макаренко Л. Обзор и оценка проблемных кредитов: потенциал рынка. 2010. 71 с. URL: http://www1.ifc.org/wps/wcm/connect/f33e340048fc6638b509bd849537832d/RussiaCR-NPL-SurveyReport-RU.pdf?MOD=AJPERES (дата звернення: 15.10.2018).

14. Лаврушин О. И. Банковский менеджмент : учебник. Москва : КНОРУС, 2009. 560 с.

15. Рабец Н. Меры по предотвращению проблемной задолженности. *Финансовый директор*. 2011. № 5. С. 54–57.

16. Слобода Л., Дунас Н. Напрями вдосконалення роботи банків України з проблемними активами в посткризовий період. *Вісник НБУ*. 2011. №4. С. 46–51.

17. Кириченко О. Банківський менеджмент : навчальний посібник для вищих навчальних закладів. Київ : ОСНОВИ, 1999. 671 с.

18. Вовк В. Я., Хмеленко О. В. Кредитування і контроль : навчальний посібник. Київ : Знання, 2008. 464 с.

19. Денисенко М. П. Кредитування та ризики : навчальний посібник.

Київ : Видавничий дім «Професіонал», 2008. 480 с.

20. Мороз А. М. Кредитний менеджмент : навчальний посібник. Київ : КНЕУ, 2009. 399 с.

21. Руководство по кредитному скорингу / под ред. Элизабет Мэйз. Перевод с англ. И.М. Тикота. Минск : Гревцов Паблишер, 2008. 464 с.

22. Siddiqi N. Credit Risk Scorecards. Developing and Implementing Intelligent Credit Scoring. New York : Wiley and Sons, 2005. 196 p.

23. Brockman M. J., Wright T.S. Statistical Motor Rating: Making Effective Use of Your Data. *Journal of Institute of Actuaries*. 1992. № 119. P. 457–543.

24. Корн Г., Корн Т. Справочник по математике: для научных работников и инженеров. Москва : Наука, 1970. 720 с.

25. Феллер В. Введение в теорию вероятностей и ее приложения. Москва : Мир, 1984. 751 с.

ДОДАТОК А. ЛІСТИНГ ПРОГРАМИ

```
library(tibble)
library(caret)
library(pROC)
library(car)

summary(AllDF)
summary(FinalDF)
plot(FinalDF$DPD, FinalDF$PaysSum/FinalDF$DaysInWork, xlab = "DPD", ylab
= "PaysSum/DaysInWork")
plot(FinalDF$BodyDelay, FinalDF$PaysSum/FinalDF$DaysInWork, xlab =
"BodyDelay", ylab = "PaysSum/DaysInWork")
plot(FinalDF$BodyBalance, FinalDF$PaysSum/FinalDF$DaysInWork, xlab =
"BodyBalance", ylab = "PaysSum/DaysInWork")
plot(FinalDF$Debt, FinalDF$PaysSum/FinalDF$DaysInWork, xlab = "Debt", ylab
= "PaysSum/DaysInWork")
plot(FinalDF$Age, FinalDF$PaysSum/FinalDF$DaysInWork, xlab = "Age", ylab
= "PaysSum/DaysInWork")
plot(FinalDF$DPD, FinalDF$PaysSum)
plot(FinalDF$DPD, FinalDF$PaysSum)

nrow(AllDF[AllDF$DaysInWork < 30, ])
sum(AllDF[AllDF$DaysInWork < 30, ]$PaysSum)

nrow(FinalDF[FinalDF$IsATO == 0, ])
sum(FinalDF[FinalDF$IsATO == 1, ]$PaysSum)

FinalDF <- FinalDF[FinalDF$DaysInWork > 30, ]

in.train <- createDataPartition(y = FinalDF$PaysSum, p = 0.75, list=FALSE)

glimpse(in.train)

FinalDF.train <- FinalDF[in.train,]
FinalDF.test <- FinalDF[-in.train,]
```

```
FinalDF.train.fit <- lm(I(PaysSum/DaysInWork) ~ . - HasPaid - HasPaidPredicted
                    + I(Debt^2)
                    #+ I(BodyBalance^2)
                    + I(BodyDelay^2)
                    + I(Debt^3)
                    #+ I(BodyBalance^3)
                    + I(BodyDelay^3)
                    , data=FinalDF.train)
summary(FinalDF.train.fit)
plot(FinalDF.train.fit)

FinalDF.train.fit$fv <- FinalDF.train.fit$fitted.values
FinalDF.train.fit$fv[FinalDF.train.fit$fv < 0] <- 0
glimpse(FinalDF.train.fit$fv)

FinalDF.train.logfit <- glm(HasPaid ~ . - PaysSum - DaysInWork -
HasPaidPredicted
                        + I(Debt^2)
                        #+ I(BodyBalance^2)
                        + I(BodyDelay^2)
                        + I(Debt^3)
                        #+ I(BodyBalance^3)
                        + I(BodyDelay^3)
                        , family = ("binomial")
                        , data = FinalDF.train)
summary(FinalDF.train.logfit)
plot(FinalDF.train.logfit)


FinalDF.train.roc <- roc(FinalDF.train$HasPaid, FinalDF.train.logfit$fitted.values,
plot = TRUE)
FinalDF.train.roc$auc

FinalDF.train.rocDF <- data.frame(sp = FinalDF.train.roc$specificities,
                          se = FinalDF.train.roc$sensitivities,
                          thr = FinalDF.train.roc$thresholds)
attach(FinalDF.train.rocDF)
```

```
FinalDF.train.rocDF[sp == max(FinalDF.train.rocDF[se>0.9, ]$sp), ]
FinalDF.train[FinalDF.train.logfit$fitted.values >= 0.03078016,
]$HasPaidPredicted <- 1
FinalDF.train[FinalDF.train.logfit$fitted.values <  0.03078016,
]$HasPaidPredicted <- 0
summary(FinalDF.train$HasPaidPredicted)

table(FinalDF.train$HasPaid, FinalDF.train$HasPaidPredicted)
#FinalDF.train$HasPaidPredicted <- FinalDF.train.logfit$fitted.values

FinalDF.train.reduced <- FinalDF.train[FinalDF.train$HasPaidPredicted == 1,]
FinalDF.train.reduced.fit <- lm(I(PaysSum/DaysInWork) ~ . - HasPaid -
HasPaidPredicted
                        + I(Debt^2)
                        #+ I(BodyBalance^2)
                        + I(BodyDelay^2)
                        + I(Debt^3)
                        #+ I(BodyBalance^3)
                        + I(BodyDelay^3)
                        , data=FinalDF.train.reduced)
summary(FinalDF.train.reduced.fit)

FinalDF.train.reduced.fit$fv <- FinalDF.train.reduced.fit$fitted.values
FinalDF.train.reduced.fit$fv[FinalDF.train.reduced.fit$fv < 0] <- 0

y1 <- predict(FinalDF.train.fit, FinalDF.test)
glimpse(y1)
y1[y1 < 0] <- 0
y1 <- y1*FinalDF.test$DaysInWork

y2.prob <- predict(FinalDF.train.fit, FinalDF.test)
glimpse(y2.prob)
summary(y2.prob)
y2.prob <- 1/(1+exp(-y2.prob))

y3 <- predict(FinalDF.train.reduced.fit, FinalDF.test)
y3 <- y3*FinalDF.test$DaysInWork
```

```
y2 <- y3
y2[y2.prob < 0.03078016] <- 0

y2[y2 < 0] <- 0
y3[y3 < 0] <- 0

comparison <- data.frame(real = FinalDF.test$PaysSum, y1 = y1, y2 = y2, y3 =
y3)

attach(comparison)
sum( (FinalDF.test$PaysSum - y1)^2 )
sum( (FinalDF.test$PaysSum - y2)^2 )
sum( (FinalDF.test$PaysSum - y3)^2 )

sum(FinalDF.test$PaysSum)
sum(y1)
sum(y2)
sum(y3)

AIC(FinalDF.train.fit, FinalDF.train.reduced.fit)

summary(FinalDF.train.fit)
summary(FinalDF.train.reduced.fit)

FinalDF.train.fit2 <- lm(I(PaysSum/DaysInWork) ~ . - HasPaid - HasPaidPredicted
- IsMale, data=FinalDF.train)
summary(FinalDF.train.fit2)
AIC(FinalDF.train.fit, FinalDF.train.fit2)

FinalDF.train.reduced.adj <- FinalDF.train.reduced
FinalDF.train.reduced.adj[FinalDF.train.reduced.adj$BodyDelay == 0,
]$BodyDelay <- 0.01
FinalDF.train.reduced.adj[FinalDF.train.reduced.adj$BodyBalance == 0,
]$BodyBalance <- 0.01

FinalDF.train.reduced.fit2 <- lm(I(PaysSum/DaysInWork) ~ Age +
                                    IsATO +
                                    log(BodyDelay) +
```

```
                                        log(BodyBalance) +
                                        log(Debt) +
                                        I(DPD^0.5),
data=FinalDF.train.reduced.adj)
summary(FinalDF.train.reduced.fit2)
AIC(FinalDF.train.reduced.fit, FinalDF.train.reduced.fit2)

FinalDF.train.adj <- FinalDF.train
FinalDF.train.adj[FinalDF.train.adj$BodyDelay == 0, ]$BodyDelay <- 0.01
FinalDF.train.adj[FinalDF.train.adj$BodyBalance == 0, ]$BodyBalance <- 0.01

FinalDF.train.fit2 <- lm(I(PaysSum/DaysInWork) ~ Age +
                         IsATO +
                         log(BodyDelay) +
                         log(BodyBalance) +
                         log(Debt) +
                         DPD, data=FinalDF.train.adj)
summary(FinalDF.train.fit2)
AIC(FinalDF.train.fit, FinalDF.train.fit2)

FinalDF.test.adj <- FinalDF.test
FinalDF.test.adj[FinalDF.test.adj$BodyDelay == 0, ]$BodyDelay <- 0.01
FinalDF.test.adj[FinalDF.test.adj$BodyBalance == 0, ]$BodyBalance <- 0.01

y4 <- predict(FinalDF.train.fit2, FinalDF.test.adj)
glimpse(y4)
y4[y4 < 0] <- 0
y4 <- y4*FinalDF.test$DaysInWork
sum(y4)

y5 <- predict(FinalDF.train.reduced.fit2, FinalDF.test.adj)
glimpse(y5)
y5[y5 < 0] <- 0
y5 <- y5*FinalDF.test$DaysInWork
sum(y5)

FinalDF.train.paid <- FinalDF.train[FinalDF.train$HasPaid == 1, ]
FinalDF.train.paid.fit <- lm(I(PaysSum/DaysInWork) ~ Age +
```

```
                    IsMale +
                    IsATO +
                    BodyDelay +
                    BodyBalance +
                    DPD, data=FinalDF.train.paid)
summary(FinalDF.train.paid.fit)

y6 <- predict(FinalDF.train.paid.fit, FinalDF.test)
y6 <- y6*FinalDF.test$DaysInWork
sum(y6)
y6[y2.prob < 0.9369124] <- 0
sum(y6)

boxTidwell(I(PaysSum/DaysInWork) ~ Age, data = FinalDF.train)
FinalDF.Age.fit <- lm(I(PaysSum/DaysInWork) ~ I(Age^3.7938), data =
FinalDF.train)
summary(FinalDF.Age.fit)

FinalDF.IsATO.fit <- lm(I(PaysSum/DaysInWork) ~ IsATO, data = FinalDF.train)
summary(FinalDF.IsATO.fit)

boxTidwell(I(PaysSum/DaysInWork) ~ DPD, data =
FinalDF.train[FinalDF.train$DPD > 0, ])
FinalDF.DPD.fit <- lm(I(PaysSum/DaysInWork) ~ I(DPD^0.5), data =
FinalDF.train)
summary(FinalDF.DPD.fit)

boxTidwell(I(PaysSum/DaysInWork) ~ Debt, data = FinalDF.train)
FinalDF.Debt.fit <- lm(I(PaysSum/DaysInWork) ~ log(Debt), data =
FinalDF.train)
summary(FinalDF.Debt.fit)

FinalDF.Age.fit <- lm(I(PaysSum/DaysInWork) ~ Age, data = FinalDF.train)
summary(FinalDF.Age.fit)
```